Journal of Artificial Intelligence Research 66 (2019) 341-380 Submitted 12/2018; published 09/2019

A Survey on Temporal Reasoning for Temporal Information Extraction from Text

Artuur Leeuwenberg TUUR.LEEUWENBERG@CS.KULEUVEN.BE
Marie-Francine Moens SIEN.MOENS@CS.KULEUVEN.BE
KU Leuven – Department of Computer Science
Celestijnenlaan 200A, 3001 Leuven,
Belgium

Abstract

Time is deeply woven into how people perceive, and communicate about the world. Almost unconsciously, we provide our language utterances with temporal cues, like verb tenses, and we can hardly produce sentences without such cues. Extracting temporal cues from text, and constructing a global temporal view about the order of described events is a major challenge of automatic natural language understanding. Temporal reasoning, the process of combining different temporal cues into a coherent temporal view, plays a central role in temporal information extraction. This article presents a comprehensive survey of the research from the past decades on temporal reasoning for automatic temporal information extraction from text, providing a case study on how combining symbolic reasoning with machine learning-based information extraction systems can improve performance. It gives a clear overview of the used methodologies for temporal reasoning, and explains how temporal reasoning can be, and has been successfully integrated into temporal information extraction systems. Based on the distillation of existing work, this survey also suggests currently unexplored research areas. We argue that the level of temporal reasoning that current systems use is still incomplete for the full task of temporal information extraction, and that a deeper understanding of how the various types of temporal information can be integrated into temporal reasoning is required to drive future research in this area.

1. Introduction

The phenomenon of time has a major influence on how people perceive, and communicate through language. Consequently, our language utterances are filled with cues about the timing of the events that we communicate about. **T**emporal **I**nformation **E**xtraction (TIE) is the process of automatically extracting temporal cues from text, with the final goal to construct a (possibly underspecified) timeline of events from them, as shown in Figure 1.

Temporal information extraction not only plays a major role in the general problem of natural language understanding (NLU), but is also used in many applications, like information retrieval (Campos, Dias, Jorge, & Jatowt, 2015), question answering (Llorens, Chambers, UzZaman, Mostafazadeh, Allen, & Pustejovsky, 2015; Höffner, Walter, Marx, Usbeck, Lehmann, & Ngonga Ngomo, 2017; Meng, Rumshisky, & Romanov, 2017; Sun, Cheng, & Qu, 2018; Pampari, Raghavan, Liang, & Peng, 2018), and multi-document summarization (Barzilay & McKeown, 2005; Ng, Chen, Kan, & Li, 2014), and has great potential in the clinical domain, for applications like patient timeline visualization (Jung, Allen, Blaylock, De Beaumont, Galescu, & Swift, 2011), forecasting treatment effects (Augusto, 2005; Zhou

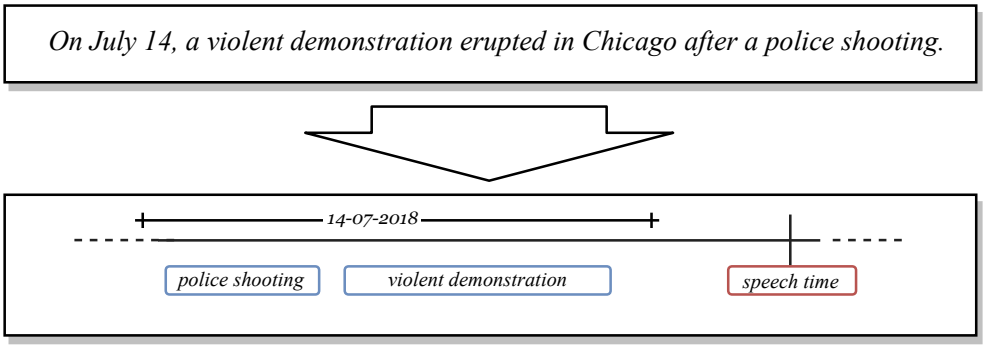

Figure 1: An example of temporal information extraction.

& Hripcsak, 2007; Choi, Bahadori, Sun, Kulas, Schuetz, & Stewart, 2016a), better early detection of diagnoses (Choi, Schuetz, Stewart, & Sun, 2016b), or patient selection for clinical trials (Raghavan, Chen, Fosler-Lussier, & Lai, 2014). Because of the strong linear structure of time itself, and the great variation in the types of temporal cues we can express in language, a central challenge in temporal information extraction is how to combine all these separate cues into a single coherent temporal ordering of the described events. To obtain this temporal ordering from many different cues temporal reasoning is of vital importance. We define **T**emporal **R**easoning (TR) in the context of TIE as the process of combining different (annotated or extracted) temporal cues to derive additional temporal information about the text. TR is crucial for TIE and has already been exploited widely in the research community in every step in the process of TIE model construction: annotation, pre-processing, model training, inference, and evaluation. Despite the importance of TIE for NLU and the crucial role of TR in TIE, there has not yet been a survey covering the research in this area.

1.1 Focus

For successful TR in practical settings two factors are very important: (1) The *expressiveness* of the TR mechanism: How complete is the temporal knowledge that the TR mechanism can infer from the temporal cues? And (2), the *efficiency* of TR: What are the computational costs of TR to infer that new temporal knowledge. These two points will get extra attention in this survey.

Although most research has focused on extraction models for certain types of temporal cues, this survey focuses on the big picture of complete TIE, where the aim is to extract all temporal cues from the text and combine them into a single coherent temporal view, for which a good TR mechanism is crucial. To cover the evolution of TR approaches for TIE in parallel with the popularization of using machine learning (ML) methods for natural language processing (NLP), the focus of this survey lies on the research on TR for TIE systems from the past three decades. We abstract from what linguistic features are successful for TIE systems, as these are discussed in depth by Derczynski (2017), and can be considered complementary to the focus of this survey.

1.2 Contributions

In the context of the interesting and important new developments in the past years this survey provides the following contributions:

- A clear explanation introducing the theory and background on TR for TIE required to comprehend the latest state-of-the-art TIE models.
- A structured overview of the various ways in which TR has been exploited in TIE models over the past three decades: in annotation, pre-processing, training, prediction, and evaluation.
- A distillation of the most important conclusions to successfully incorporate TR in a TIE system.
- Directions for future work and on promising unexplored avenues in the research area of TIE.

1.3 Structure

The survey is structured as follows: First, in section 2, we provide an exemplified overview of what types of temporal information are present in natural language texts. These include relative and absolute cues, definite and indefinite cues, implicitness of temporal information, world knowledge, and the role of under-specification, as these aspects play an important role in temporal reasoning. In section 3, we introduce the theory of temporal reasoning that is required to comprehend and assess the current state-of-the-art models and methods, as discussed in the following sections. Section 4 describes the most widely used annotation schemes for temporal information, and discusses how they relate to temporal reasoning frameworks. In section 5, we arrive at the core of this survey, and provide a complete and comprehensive overview of the literature on TR for TIE. Then, in section 6, we give suggestions on promising directions and less explored areas based on the earlier sections. Lastly, in section 7, we summarize the most important findings and conclusions of the survey.

2. Temporal Information in Language

In this section, we give a short (exemplified[1]) overview of different temporal cues that can be expressed in language to show what types of temporal information the cues can provide, i.e., in what way the cues may possibly constrain the position of event intervals on the timeline. We focus less on the different ways temporal cues can be expressed, i.e., linguistic variation, as this has no direct impact on TR.

It is important to study the types of temporal information that can be expressed by temporal cues because the different cues need to be combined by a TR system in order to build a complete temporal picture of the text, or construct a timeline.

1. Examples from the New York Times, and the clinical i2b2 corpus (Sun, Rumshisky, & Uzuner, 2013).

2.1 Timeline Components Captured by Temporal Cues

Temporal cues can bound various components of the event timeline: full positions of *intervals*, but also just the *start*, *end*, or *duration* of intervals.

For instance, in Example 1, the duration of the antibiotics administration is given (10 days), and so is its start time (somewhere on the 2nd of June). While, for the improvement of the respiratory status only the end time is mentioned explicitly (last 2-3 days of the antibiotics administration). This shows that for a fairly simple text fragment, a TR system already needs to be able to combine temporal bounds on at least three different components of the timeline.

Example 1. *Antibiotics were started on 6/2 and continued for 10 days. Respiratory status improved up til the last 2-3 days.*

2.2 Relative and Absolute Cues

As shown in the previous example, temporal cues can provide *absolute* references to the timeline, by referring to absolute intervals, like dates or times, or absolute durations, like a certain number of hours. In Example 2, the duration of the third set (*28 minutes*) is an absolute cue. However, additionally quite often the temporal cues provide *relative* information. In the example, there are three explicit relative cues: (1) the duration of the first two sets are *less than* 1 hour, (2) the third set started *after* the first two, And (3) the whole situation took place in the past, i.e., before the speech time (ST) of the sentence, indicated by the past tense of the verbs. The ideal TR system should be able to resolve combinations of absolute and relative cues.

Example 2. *After the grueling duels of the first two sets, which each had taken nearly an hour, Nadal won the third set in 28 minutes.*

2.3 Definite and Indefinite Cues

Quantification of timeline components can be definite, referring to clear quantities, or indefinite, using vague quantification.

In Example 3, the duration of the patient being HIV positive is definite (*2 years*). Whereas the duration of the left upper quadrant pain is, although explicit, quantified vaguely (*long-standing*), and hence indefinite.

Example 3. *The patient is a 27-year-old woman who is HIV positive for two years. She presented with left upper quadrant pain which is a long-standing complaint.*

2.4 Implicitness and World Knowledge

A significant part of the temporal information conveyed in text is implicit. Event durations, and event position are often implicit, or considered common world knowledge. In Example 4, although not mentioned explicitly the *charges* have been made before the judge's question. Also, the man's answer probably lasted only a few seconds, and happened clearly after the judge's question. Whereas, if we would have replaced *answered* with *trembled*, this would probably have lasted longer, and was possibly already going on during the judge's question.

Example 4. *Wearing a black scarf, slim-fitting navy suit and tortoiseshell glasses, he said little and answered, "I do, your honor", when the judge asked if he understood the charges.*

2.5 Underspecification

A major aspect of temporal information extraction is underspecification. Almost in all cases, temporal cues do not provide a fully specified timeline (full absolute positioning of events on the calendar), leaving open multiple temporal interpretations. As can be seen in all previous examples 1-4, it was never mentioned, for example, in what year the events took place, allowing multiple valid timeline interpretations, something that temporal reasoning systems should be able to deal with appropriately.

3. Frameworks for Temporal Reasoning

In the previous section we observed that language can contain many different types of temporal information. To combine all these different types of temporal information into a coherent temporal view, or timeline, we require temporal reasoning. To reason with different types of temporal information about events, several frameworks have been developed. The temporal interpretation of an event is generally considered as an interval on the timeline. The span of the interval corresponds to the time that the event takes place. Consequently, TR often addresses reasoning with intervals. As a reference, Fisher et al. (2005) provide an overview of general TR, but do not discuss TIE systems in detail. Here, we review the TR frameworks that have been used for TIE, as a back-bone for section 5, where the integration of TR in TIE systems will be discussed.

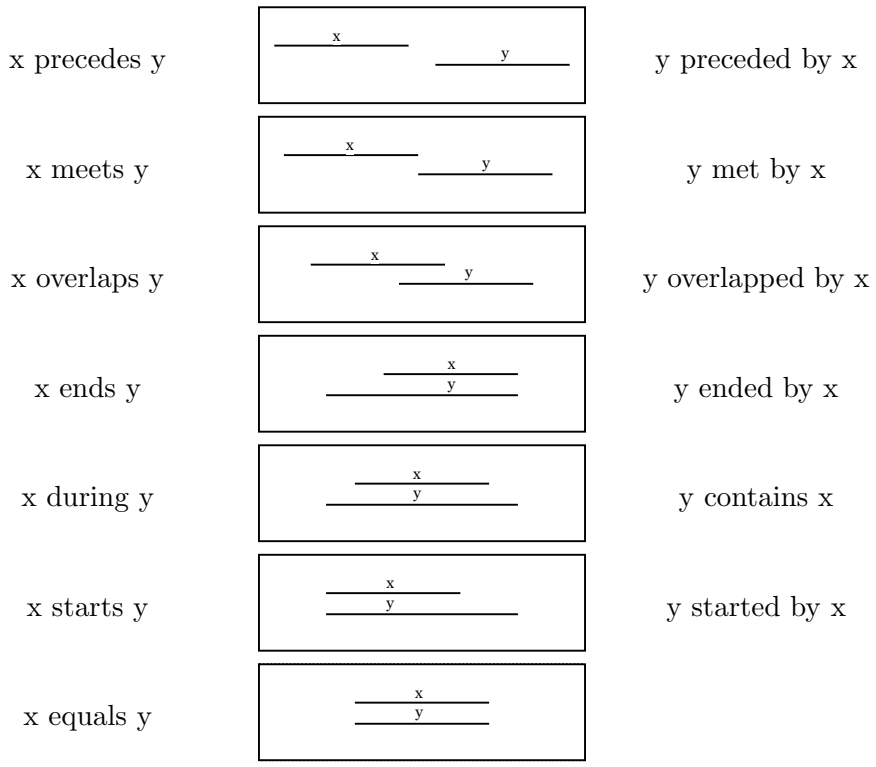

Figure 2: Allen's thirteen basic interval relations.

3.1 Allen Interval Relations

One of the most popular TR frameworks used in TIE was proposed by Allen (1983). He proposed a set of thirteen mutually exclusive *basic* interval relations that could be assigned to any pair of definite intervals. These relations and the corresponding operations are known as Allen's interval algebra. All thirteen basic relations and their visualizations are shown in Figure 2. As can be seen, the thirteen basic relations are six pairs of converse relation pairs, and the *equals* relation, which is symmetric. From these basic relations that can only represent relations between *definite* intervals, where relative positions of start and end-points are known, indefinite interval relations can be constructed, where the start or end of the intervals might be incomplete. In Allen's algebra each indefinite relation (called a *general* Allen relation) is represented as a *disjunctive* set of basic relations. The set representation of a general Allen relation between two events contains all basic relations that are possible between the events, given the (incomplete) information about their starts and endings.

Example 5. *y started sometime during x*

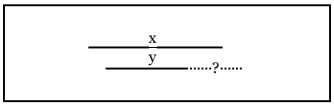

Figure 3: Visualization of Example 5 with the general Allen relation: x {contains, overlaps, ended by} y.

For example, the sentence of Example 5 could be represented by Figure 3. Only relative information about the start of y is known, namely that it lies within the boundaries of x. However, there is no information given about the end of y, making it impossible to assign a basic Allen relation, as the intervals are indefinite. The correct general Allen relation is: x {*contains*, *overlaps*, *ended by*} y, and indicates that any of these three basic relations in the set, *contains*, *overlap*, or *ended by*, could apply to the situation. The full set of Allen interval relations is the power set of the basic relations, resulting in $2^{13} = 8192$ relations. Notice that when no information about the relation between two intervals (or events) x and y is given, any basic relation is possible. So, in that case the general Allen relation between x and y would be represented by the set containing all basic relations. In other words, the less we know, the more is possible, so the bigger the representation.

3.1.1 Temporal Closure

To infer new relations from a set of general Allen relations relating different events, a composition table is used. The table contains transitivity rules for all basic relations, i.e., it shows for any pair of connected relations r_1(x, y) and r_2(y, z), what relation r_3(x, z) could be inferred. An example for the transitivity for the *precedes* relation is given in Figure 4. Using this principle, a temporal closure (called 'Propagate' in the original paper) can be calculated, adding new relations to the existing set. Computing the full closure, which includes all possible inferences that can be made, is NP-complete, making it highly intractable (Vilain, Kautz, & Van Beek, 1990). Often, in practice, only a subset of the transitivity rules is used.

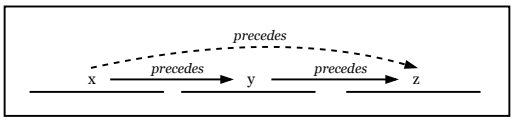

Figure 4: An example of an inferable precedes relation (dashed) through transitivity: x precedes z is inferred from the fact that x precedes y and y precedes z.

3.1.2 Temporal Consistency

An important task in TR is checking temporal consistency for a set of relations, as a timeline can only be constructed from a consistent set of relations. In Allen's algebra, temporal consistency can be calculated by checking if, when going through all possible chains of inference, the intersection of all inferable relations for each pair is not empty. In other words, an assignment of general Allen relations is consistent if at least one basic Allen relation can be assigned to each pair of intervals, after closure.

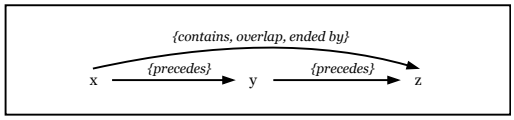

Figure 5: An example of an inconsistent assignment of Allen relations.

An example of an inconsistent assignment of relations is shown in Figure 5. The example is inconsistent because from the fact that x *precedes* y and y *precedes* z it can be inferred that x {*precedes*} z. And, when taking the intersection for pair (x,z) of the inferred relation x {*precedes*} z and the already assigned relation x {*contains*, *overlap*, *ended by*} z we end up with the empty set. This indicates there are no possible basic relations for pair (x,z). From this we can conclude that the example is not consistent. Calculating temporal consistency is, like the temporal closure, NP complete when using the full Allen algebra (Vilain et al., 1990) as it requires temporal closure. This high computational complexity is not very practical in real applications. For this reason, more efficient solutions have been proposed, which we will cover in the next section.

3.2 Subfragments of the Full Allen Algebra and Point Temporal Algebra

Because of the high complexity of calculating temporal closure and consistency for the full Allen algebra, many different more tractable sub-fragments of Allen's interval algebra have been proposed (Vilain et al., 1990; Freksa, 1992a; Nebel & Bürckert, 1995; Ligozat, 1996; Krokhin, Jeavons, & Jonsson, 2003). Some also focus on integrating quantitative reasoning (Dechter, Meiri, & Pearl, 1991; Meiri, 1996; Dechter & Cohen, 2003) or uncertainty (Schockaert & De Cock, 2008). Although most current research in TIE systems has focused on using Allen relations, representing mostly relative interval cues, the ability to combine quantitative temporal cues and being able to deal with uncertainty is very important, as the variance of cues in language is vast, as seen in the previous section. And all cues need to be taken into account to construct a fully coherent temporal view from the text. We will not discuss all the sub-fragments and extensions in this survey as the vast majority of these extensions have not been used in current TIE systems. Rather, we provide a theoretical back-bone on which many of these sub-fragments are built. We also use this back-bone to define a categorization of TR-expressiveness in which we will later classify the different TR approaches used for current TIE systems.

A major insight on which many efficient TR algorithms are built is the fact that the basic Allen interval relations can be expressed as sets of point-relations in the *point temporal algebra*. From this perspective, each pair of intervals (x,y) can be seen as a set of four points: the starts of both intervals x^-, and y^-, and their endings: x^+, and y^+. In the point temporal algebra, there are three point-wise relations that can occur between each of these points: $<$, $=$, and $>$. These point-relations can be used to express each basic Allen relation as a *conjunctive* set of point relations on the start and endings of the intervals. For example, x {*equals*} y can be expressed by the conjunctive set $\{x^- = y^-, x^+ = y^+\}$, i.e., iff the start and endings of intervals x and y are equal, then x and y are also equal. As another example, x{*precedes*}y can be expressed as $\{x^+ < y^-\}$, i.e., iff the end of x lies before the start of y then interval x lies before interval y. In general, because the four points describe starts and

endings of intervals, it is always the case that each interval's start x^- lies before its end x^+.

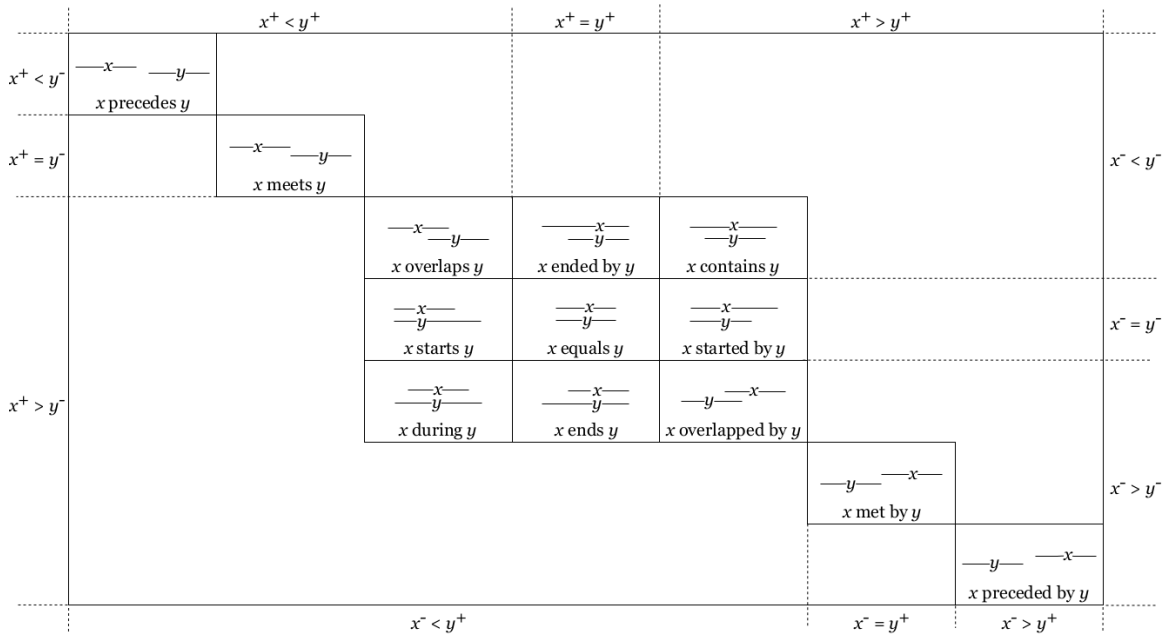

Figure 6: The lattice showing the relation between point algebra and the basic Allen interval relations, and the conceptual neighborhood between interval relations (Freksa, 1992a).

Based on their point algebraic representations, Allen's basic interval relations can be ordered in a very informative lattice (Freksa, 1992a), as shown in Figure 6. From this lattice we can read several things:

1. How to convert a basic Allen interval relation to a set of point algebraic constraints: start from a basic interval relation box in the lattice, and then take the union of all point-algebraic constraints on the corresponding outsides of the rectangle.

2. How to construct a (general) Allen relation from a set of point-algebraic constraints: start from the point-algebraic constraints on the outside of the rectangle, and take the intersection of the interval relations that are covered by the area inside the rectangle corresponding to the point-algebraic constraints.

3. How to determine the conceptual neighborhood between two basic Allen interval relations: count how many boxes have to be crossed to get from one basic interval relation to the other (i.e., how much do the end-points need to shift to change from one interval relation to the other).

For example, if we have a cue saying that the start of event y happens somewhere during event x, i.e., $\{x^- < y^-, x^+ > y^-\}$ (as in Example 5), we can read from the lattice what relations are covered by the overlapping area of these constraints: {*contains*, *overlaps*, *ended by*}, from which we can conclude that the corresponding general Allen relation is x {*contains*, *overlaps*, *ended by*} y.

Using this link between interval and point relations TR about intervals can be done in the point algebra, which is much more efficient, as it has only three basic relation types, resulting in a much smaller composition table. Additionally, expressing interval relations by conjunctive sets of point relations, instead of disjunctive sets of basic interval relations, ensures that when we have little temporal knowledge, the set representation is smaller, instead of bigger (as with general Allen relations). These two components contribute to the fact that TR in temporal point algebra fragments has only polynomial complexity instead of the NP-completeness of the full Allen algebra (Vilain et al., 1990). This gained efficiency and flexibility in representation are very important when considering practical TIE systems. Even more when combining temporal cues from different documents, or different sources for which complex temporal resolution of the different cues is required to obtain a coherent timeline.

As mentioned in the beginning, we cannot express each general Allen relation as sets of point-algebraic constraints, but only a fragment of them. Point algebra can only express interval relations that can be represented as a conjunction of point-algebraic constraints. For example, the general Allen relation x {*precedes*, *preceded by*} y cannot be expressed as conjunction of point-algebraic constraints. We can see this from the lattice, as we cannot capture these two interval relations in a single rectangle, without including other basic relations as well.

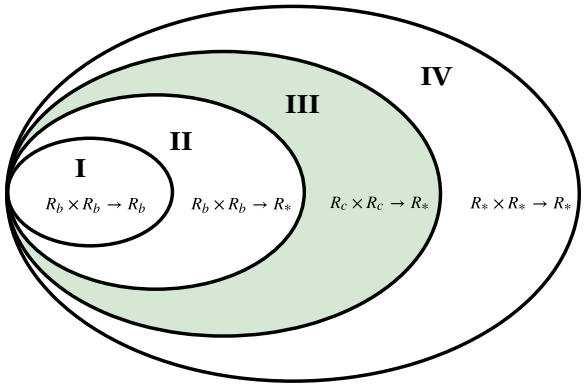

Figure 7: Onion diagram showing four classes of temporal reasoning rules, indicating expressiveness of the temporal reasoning. R_b stands for any basic Allen relation, R_c stands for any Allen relation that can be expressed as conjunction of point-algebraic constraints, and R_* can be any Allen relation. Layer III is the most expressive class that still operates in polynomial time, which is important for practical systems.

3.2.1 Expressiveness of Temporal Reasoning

Later on, in section 5, we review the currently existing TR-based TIE systems and categorize them in various ways in order to compare them. To categorize the TIE models with regard to the expressiveness of their TR component we construct four classes based on the part of the transitivity table that is used for TR. This categorization is shown in Figure 7: The most inner layer covers rules that only use basic Allen relations (R_b) in the condition and conclusion of the rules. The second layer covers TR rules that have only basic relations in

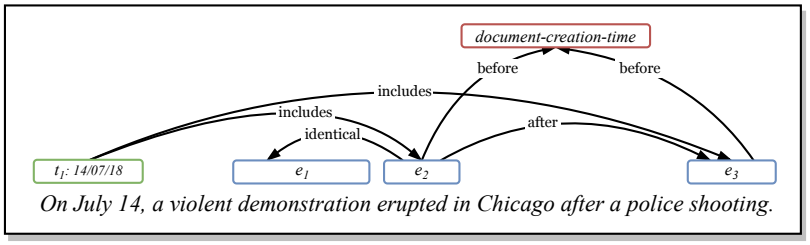

Figure 8: An example of TimeML-style annotation: Events in blue, a (normalized) temporal expression in green, the document-creation-time in red, and arrows indicating the temporal links (TLinks) among them.

their condition, but can have general Allen relations (R_*) in the conclusion. The third layer covers the sub-fragment that is translatable to point algebra, covering all rules that have relations in the condition of the rule that can be expressed as conjunction of point-algebraic constraints (R_c). Finally, the outer layer is the full Allen algebra, covering the full transitivity table. Outer layers include inner layers in terms of rule sets, and are more expressive, but TR can be more computationally complex (depending on the implementation).

Layer I type reasoning is used quite commonly, as most systems build on basic Allen relations, layer II is more expressive but used less, as it involves reasoning with sets of basic relations instead of only basic relations.

For practical TIE systems, layer III is very important for three reasons: (1) It is the most expressive fragment that can be implemented in polynomial time; (2) The full fragment can be mapped to point-algebraic constraints and back, introducing flexibility with regard to combining different types of temporal cues; and (3) It covers all basic Allen relations, which is beneficial as these are often available in the annotated data. Before discussing the TR used in TIE systems (section 5), we will discuss the annotation of temporal information.

4. Annotation of Temporal Information

The annotation scheme for temporal annotation in text steers the types of temporal cues that can be extracted by TIE systems, and therefore also how TR can be exploited. The most widely used scheme is TimeML (Pustejovsky, Castano, Ingria, Sauri, Gaizauskas, Setzer, Katz, & Radev, 2003a), which is an ISO-standard for annotating temporal information in text (Pustejovsky, Lee, Bunt, & Romary, 2010). Most temporal corpora have been annotated with TimeML, or a very similar (sub)scheme (Pustejovsky, Hanks, Sauri, See, Gaizauskas, Setzer, Radev, Sundheim, Day, Ferro, et al., 2003b; Sun et al., 2013; Cassidy, McDowell, Chambers, & Bethard, 2014; Styler IV, Bethard, Finan, Palmer, Pradhan, de Groen, Erickson, Miller, Lin, Savova, et al., 2014). And TimeML has been used in TempEval, a series of shared tasks on evaluating temporal information extraction that resulted in many of the existing TIE systems (Verhagen, Gaizauskas, Schilder, Hepple, Katz, & Pustejovsky, 2007; Verhagen, Sauri, Caselli, & Pustejovsky, 2010; UzZaman, Llorens, Derczynski, Allen, Verhagen, & Pustejovsky, 2013; Minard, Speranza, Agirre, Aldabe, van Erp, Magnini, Rigau, & Urizar, 2015; Llorens et al., 2015). We will discuss the shared consequences for TR of the TimeML scheme. In TimeML the core concepts are *event ex-*

Table 1: Temporal link (TLink) types from TimeML and their corresponding basic Allen interval relations.

TimeML TLink	Basic Allen Relation
BEFORE	*precedes*
AFTER	*preceded by*
INCLUDES	*contains*
IS INCLUDED	*during*
IMMEDIATELY BEFORE	*meets*
IMMEDIATELY AFTER	*met by*
BEGINS	*starts*
BEGUN BY	*started by*
ENDS	*ends*
ENDED BY	*ended by*
DURING	*during* | *equals*
DURING_INV	*contains* | *equals*
-	*overlap*
-	*overlapped by*
SIMULTANEOUS	*equals*
IDENTICAL	*equals*

pressions and *temporal expressions*. Event expressions refer to events in the real world, and can be of different types, like states (e.g., *bankrupt*), actions (e.g., *sailing*), occurrences (e.g., *meeting*), reporting events (e.g., *said*) among other types. Temporal expressions (timex) refer to calendar dates (e.g., *21st of August, 2018*), times (e.g., *1 o'clock*), definite durations (e.g., *two hours*), or sets of times (e.g., *3 times a week*). The function of these timex expressions is to anchor events to the calendar timeline. In TimeML, events and timex expressions are temporally connected through temporal links (TLinks). TLinks can have thirteen types that almost (but not exactly) follow Allen's basic interval relations. An example of a TimeML-annotated sentence in shown in Figure 8. The TLink types are shown in Table 1. TLinks can be annotated between three categories of candidate pairs:

1. Between events (EE-R).
2. Between temporal expressions and events (TE-R)
3. Between each event and the document-creation-time (DCT-R).

As can be seen in the Table most TLinks match with a basic Allen relation. However, Allen's *overlap*, and *overlapped by* relations are not represented (UzZaman & Allen, 2011), and the temporal interpretation of DURING and DURING_INV relations seem similar to IS INCLUDED and INCLUDES, but are not clearly defined (Chambers, Wang, & Jurafsky, 2007; Derczynski, Llorens, & UzZaman, 2013; Derczynski, 2016), and are also sometimes interpreted as SIMULTANEOUS (UzZaman et al., 2013). The difference between SIMULTANEOUS and IDENTICAL is that SIMULTANEOUS can apply to two different events happening at the same time, whereas IDENTICAL means two event mentions refer to the exact same event (event co-reference).

In terms of expressiveness, TimeML models a small subset of the full Allen algebra (2^{13} relations), and sticks to modeling the basic Allen relations. This is because temporal annotation is a complex task for annotators, and annotation complexity needs to be taken into account to obtain high-quality annotations with reasonable inter-annotator agreement.

Nevertheless, the expressiveness of TimeML is expanded by also including timex annotations as calendar anchors. Because timexes of types *date* and *time* can be temporally interpreted as absolute intervals with clear positions on the calendar that carry a clear temporal ordering with respect to each other (e.g., *1990* is always before *1991*). And similarly, timex with type *duration* can be interpreted as quantified interval durations that come with an implicit order on durations (e.g., *1 hour* is always shorter than *2 hours*). Hence, there is a fourth category of temporal links that can be automatically derived from timex annotations:

4 TLinks between temporal expressions (TT-R).

While using timexes as calendar anchors to increase the amount of temporal information captured, the expressivity of TimeML is limited by the expressivity of basic Allen relations. As shown in sections 2 and 3, the temporal information that can be expressed in language not always concerns definite intervals, for which basic Allen relations are ideal. Underspecification of an event's duration or ending in the text could potentially cause disagreement between annotators, as no basic Allen relation would be suitable. However, including general Allen relations into the annotation scheme, to model temporal uncertainty, could make the annotation task very complex.

As in the construction of many TimeML corpora annotators are not forced to annotate all candidate pairs, this regularly results in sparse TLink annotations. Sparse annotations can (1) make extraction difficult because of class imbalance, and (2) cause problems in evaluation because extraction systems can get penalized for predicting relations that the annotator may have missed. An attempt to address sparse temporal graphs is the TimeBank Dense corpus by Cassidy et al. (2014), who explicitly asked annotators to annotate relations between all events, within a certain token window. Whenever no basic Allen relation could be assigned, annotators are asked to assign the label *vague*. Although *vague* does not tell a lot about the degree of temporal uncertainty, and could in fact be replaced with the general Allen relation including all basic relations in terms of reasoning, it does make very clear what pairs have clear orderings, and what pairs do not. An interesting unexplored avenue might be to annotate the *vague* relations in the TimeBank Dense with their general Allen relations to obtain an even more complete temporal graph.

Ning et al. (2018c) recently addressed the issue of temporal uncertainty as well, and also pose the question whether all events can actually be related to each other temporally. They propose a multi-axis annotation scheme where they first separate events that are anchorable on the timeline, from other events (negated events, opinions, intentions). They assume that if two events are on the same axis, they can be temporally related. In a second step, they ask annotators to annotate temporal point-wise relations (*before*, *after*, *equals*, and *vague*) between the starting points of the anchorable events, instead of the conventional interval relations, obtaining high inter-annotator agreement even on crowd-sourced annotations. They found when also asking annotators to annotate relations between end-points this was found much more difficult by the annotators. This is possibly explained by the difference

in interpretation of event durations between annotators, as these are often not explicitly mentioned in the text, and assume background knowledge shared by the readers and writers. This scheme is interesting as it annotates a different set of relations than the ones used by TimeML. We can see the relation between the two clearly from the lattice in Figure 6, as the relations between start points can be separated by the right y-axis of the lattice ($x^- < y^-$, $x^- < y^-$, $x^- > y^-$). This shows that if a TR component would be able to deal with both interval and point-relations, data from both schemes could be combined this way to learn or evaluate TIE systems, which could be very useful.

Although Ning et al. (2018c) do not describe duration annotations, TimeML includes duration annotations. In TimeML, cues on duration need to be explicitly mentioned by a timex in the text in order to be annotated. The fact that a *long meeting* takes longer than a *short meeting* is not annotated as *long* and *short* are not timex expressions. Also background knowledge about typical event duration, in case the duration is not mentioned explicitely by a timex, is not annotated by TimeML. Pan et al. (2006a, 2006b, 2011) proposed an annotations scheme fully dedicated to annotating event durations. They asked annotators to provide quantified bounds on the durations of the events mentioned in the texts (like *1-10 minutes*, or *1-2 days*). This annotation is done on the event level and does not explicitely annotate timex expressions. This way annotators are free to use any cues they can find in the text, or their background knowledge[2].

Another recent scheme including event durations was proposed by Reimers et al. (2016, 2018). They classify events into two coarse duration types: one-day and multi-day events. Similarly to TimeML they annotate timex expressions and link events to these timex expressions when possible. However, when this is not possible, annotators are asked to directly provide calendar bounds on when each event happened (like *after 1992*, and *before 2000*). This way, the events can also be temporally related to calendar anchors that do not occur directly in the text as timex expressions, in contrast to the schemes mentioned earlier. Because this scheme annotates on the event level, and not on the event-pair level, its annotation time is linear. Also, when all events are linked to absolute calendar dates many TLinks can be automatically inferred, by exploiting the order on calendar days. This way a lot of temporal information can be captured by relatively little annotation. Although minor limitations of this scheme are the coarse granularity of the durations, and the fact that now relative ordering statements between events cannot be annotated directly, we believe the direction of annotating on the event level, and allowing non-explicit timeline anchors to be very promising.

We can see that in Allen algebra (and also in point-algebra) it is not possible to directly include quantitative statements about interval durations, even though people clearly express duration information, and have access to it through background knowledge to make temporal inferences. This makes it difficult to combine the duration datasets with the currently interval-based reasoning methods. Combining relative position information, and quantitative duration information during reasoning has - to our knowledge - not been done in the TIE systems in the literature and could be very interesting for future research.

2. To obtain this background knowledge on event durations automatically, instead of having to annotate it explicitly, there has been work on extracting typical event durations from the web using lexico-syntactic patterns (Kozareva & Hovy, 2011; Williams & Katz, 2012).

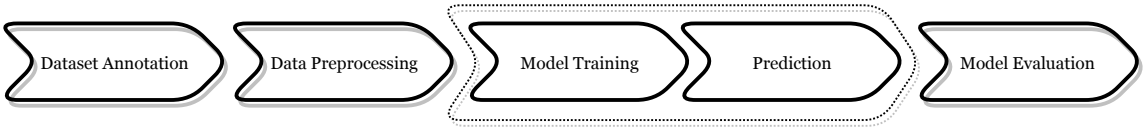

Figure 9: Steps for constructing temporal information extraction models. During each step temporal reasoning can be exploited.

5. Temporal Reasoning for Temporal Information Extraction

In this section we will review the development of TIE systems, focusing on approaches that use some form of TR. We do not discuss extraction of events as generally this does not impact TR directly. Extensive research was done on event extraction and timex extraction resulting in strong extraction systems for these tasks (Strötgen & Gertz, 2010; Chang & Manning, 2012; Derczynski, Llorens, & Saquete, 2012; Lee, Artzi, Dodge, & Zettlemoyer, 2014; Strötgen & Gertz, 2015; Miller, Bethard, Dligach, Lin, & Savova, 2015; Bethard & Parker, 2016; Strötgen & Gertz, 2016; Derczynski, Strötgen, Maynard, Greenwood, & Jung, 2016; Laparra, Xu, Elsayed, Bethard, & Palmer, 2018b; Laparra, Xu, & Bethard, 2018a; Olex, Maffey, Morgan, & McInnes, 2018). The normalization of timex, determining the calendar value of relative temporal expressions (e.g., *last year* $\rightarrow$ 2018), does involve TR. Almost all state-of-the-art systems resolve timex normalization successfully using hand-crafted rules based on lexical patterns (Mani & Wilson, 2000; Negri & Marseglia, 2004; Verhagen & Pustejovsky, 2008; Strötgen & Gertz, 2010; Kolomiyets & Moens, 2010; Chang & Manning, 2012; Llorens, Derczynski, Gaizauskas, & Saquete, 2012; Lin, Chen, & Brown, 2013; Filannino, Brown, & Nenadic, 2013; Sun, Rumshisky, & Uzuner, 2015; Mirza, 2015; Strötgen & Gertz, 2016; Derczynski et al., 2016; Real, Rademaker, Chalub, & de Paiva, 2018; Olex et al., 2018). We would like to highlight the scheme by Bethard and Parker (2016) for its explicit use of a more general TR method, called SCATE (Semantically Compositional Annotation Scheme for Temporal Expressions), in which the temporal value (interval or duration) of a timex is composed from the individual words of the timex through interval operations. How the words are to be composed is annotated in the corpus through links, which can also be predicted by a relation extraction model (Laparra et al., 2018b, 2018a; Olex et al., 2018).

For the rest of this section we focus on the task of ordering events, which builds on top of event and timex extraction and normalization, and often involves extensive TR, as information on multiple events and temporal expressions has to be combined. For this task the state-of-the-art event ordering systems still perform below application level (Bethard, Derczynski, Savova, Pustejovsky, & Verhagen, 2015; Bethard, Savova, Chen, Derczynski, Pustejovsky, & Verhagen, 2016; Bethard, Savova, Palmer, & Pustejovsky, 2017; Ning, Feng, Wu, & Roth, 2018a; Meng & Rumshisky, 2018). We will discuss the different ways in which TR can be exploited in the different steps of TIE model construction, shown in Figure 9, starting with annotation.

5.1 Temporal Reasoning for Dataset Annotation (TR-DA)

A widely recognized problem in temporal annotation, is the fact that because of the complex nature of temporal information, annotators tend to miss temporal cues in the text, especially when annotating TLinks between pairs of events (Mani, Verhagen, Wellner, Lee, & Pustejovsky, 2006). Verhagen (2005), and Setzer et al. (2005) argue that when annotating temporal relations, not all event-pairs are equally useful to annotate, as some can be deduced from the already annotated TLinks. To save annotation effort they suggest to incorporate TR in the annotation tool, and only ask annotators to label event pairs for which it is not yet possible to infer a relation from the annotations already present. Another motivation to exploit TR during annotation is that annotators sometimes provide temporally inconsistent labelings, which makes it harder to train TIE systems (Gennari & Vittorini, 2016, 2017). Verhagen et al. (2006) used TR with basic Allen relations in their TANGO toolkit. Also, in the construction of the TimeBank-Dense (Cassidy et al., 2014), where annotators are forced to annotate temporal links between all entity pairs within a certain window, after each annotation a temporal closure is calculated. This way, the annotated corpus obtained interesting properties: (1) its graphs are strongly connected, (2) the resulting corpus is consistent, and (3) all required edges are labeled. As not all events can be related through basic Allen relations, the authors introduced a *vague* relation type for when two events are difficult to relate. These works show that TR can be very beneficial during data annotation, making the use of TR for dataset annotation a very interesting direction of research needed to construct better application-level TIE systems. Also the combination of TR with active learning (Settles, 2012) as a way to collect informative training examples might be very interesting, as temporal annotation is generally considered difficult and consequently time consuming (Boguraev & Ando, 2007; Tissot, Roberts, Derczynski, Gorrell, & Del Fabro, 2015; Derczynski, 2016; Gennari & Vittorini, 2017).

5.2 Temporal Reasoning for Data Preprocessing (TR-DP)

Once a dataset has been annotated, a commonly used method is to expand the training data, as annotators frequently miss TLinks. This can be done by applying a transitive closure to infer new TLink annotations based on the ones already present. The firstly used algorithm to calculate a temporal closure for this purpose is called *SputLink* (Verhagen, 2004). SputLink is based on a subset of Allen's interval algebra and uses a composition table of 745 axioms (or transitivity rules) to infer new relations. It was used already in one of the first machine learning approaches to temporal relation extraction by Mani et al. (2006). In their experiments, using SputLink increases the number of TLinks by a factor 11. The effect of temporal closure on performance for models that use no TR during prediction is ambivalent. There have been cases where temporal closure increases performance (Mani et al., 2006; Mani, Wellner, Verhagen, & Pustejovsky, 2007), but also where it does not (Chambers & Jurafsky, 2008b), or even decreases performance (Tatu & Srikanth, 2008). When using a model that exploits TR during prediction it has been shown that data expansion through a temporal closure can be very beneficial (Chambers & Jurafsky, 2008a). Models that use TR during prediction will be described in the next section.

5.3 Temporal Reasoning for Training or Prediction (TR-TP)

To enhance temporal relation extraction models, TR can be integrated during training, and during prediction. Since training and prediction are so closely related, and often prediction is a sub-procedure of training, we discuss the integration of TR in training and prediction together in a single section.

There are various reasons why integration of TR during training and prediction can improve TIE models: (1) ensure temporal consistency among the predicted relations, and (2) to constrain the output space and improve prediction accuracy.

Predicting consistent temporal information is very important for real applications, as it is not possible to construct a proper timeline from a set of inconsistent TLinks. To ensure consistency of TLinks, the expressiveness of TR is crucial. Computational efficiency of TR is also very important for real applications, especially when the TR needs to be performed for each prediction. So, the ideal TR method for a TIE application has a good trade-off between expressiveness and computational efficiency. Many methods have been explored for incorporating TR during model training and prediction, which we discuss one by one.

5.3.1 Best-First (BF) or Natural Reading Order (NRO) Greedy Inference

One of the first approaches to incorporate TR during prediction was proposed by Bramsen et al. (2006a). They addressed the task of temporally ordering text segments in clinical narratives by first detecting segment boundaries, and afterwards classifying each pair of segments into one of three temporal relations: BEFORE, AFTER, or INCOMPARABLE. To ensure that the finally predicted temporal graph for each document is consistent, they employ a *best-first* (BF) greedy strategy: (1) All segment pairs are sorted by model confidence (based on the score predicted by a pairwise relation classifier), (2) In order of high-to-low confidence, relations are added one-by-one. After adding each relation a temporal closure is applied to expand the graph, until all pairs are connected.

A slight alteration is to order the relations not by model confidence in step 1, but by *natural reading order* (NRO). Afterwards, in step 2, the same strategy is applied as with BF: adding relations one-by-one followed by a transitive closure. Regarding performance, both NRO and BF outperform the baseline models that use no TR during prediction at all, and BF obtains the best results (Bramsen, Deshpande, Lee, & Barzilay, 2006b).

5.3.2 Post-Hoc (P-HOC) Conflict Resolution

Many approaches first use greedy prediction to predict a temporal graph, and resolve conflicts post-hoc. These approaches involve the removal of conflict-causing edges, based on model confidence (Verhagen & Pustejovsky, 2008; Tatu & Srikanth, 2008; Cheng, Anick, Hong, & Xue, 2013; Sun, 2014; Meng et al., 2017). Verhagen and Pustejovsky (2008) used different models to predict different parts of the temporal graph, and assign different confidences per model to remove the least confident conflict-causing TLinks. Cheng et al. (2013) resolve conflicts by removing the least confident conflict-causing edge for each within-sentence triangle of relations. Meng et al. (2017) remove conflict-causing TLinks such that the sum of their confidences is as low as possible. Although most works using post-hoc conflict resolution report a positive impact of the resolution, they have not been

compared to each other in a quantitative manner. This could be an interesting comparison for future research.

5.3.3 Sieve-Level Inference (SLI) and Stacked Inference (SI)

A popular method for TIE is the sieve-based method (Chambers, Cassidy, McDowell, & Bethard, 2014), as it is a flexible way to incorporate both rule-based and machine learning components. The idea of the sieve-based approach is that TLinks are extracted in different consecutive phases by different model components (or sieves). Each sieve extracts TLinks, using the original input text, and the outputs from earlier sieves. TR is incorporated into sieve-based TIE systems by taking a transitive closure on the extracted TLinks after applying each sieve. This way later sieves are prevented from assigning TLinks that are inconsistent with those extracted by earlier sieves. Another way to look at this is that the closure in fact makes the output space smaller after each sieve. Typically the sieves are ordered by precision, making the *sieve-level inference* (SLI) similar to the greedy BF approach, except relations that are not added one-by-one, but in groups, i.e., sieve-by-sieve. Very similar approaches have been adopted by Mirza and Tonelli (2016) and McDowell et al. (2017). Mirza and Tonelli (2016) separately evaluated the effect of this reasoning approach, and report an increase in recall contributing to an increase in performance when including the sieve-closure.

Another approach that exploits TR in multiple prediction stages is the *stacked inference* (SI) approach of Laokulrat et al. (2014, 2015). In the first stage their model predicts TLinks using pairwise local logistic regression classifiers, as many approaches do. However, on top of that they apply a transitive closure. Then, in a second stage, they learn a new TLink classification model that additionally takes the predicted relations from stage 1, and their corresponding probabilities as input features. An example of a stage 2 feature used to predict a TLink for some given entity pair is the set of all phase-1 TLink-paths connecting the two events. This way their model can learn to predict TLinks in context of other TLinks, resulting in a learned inference procedure.

A recent similar SI approach was used by Meng and Rumshisky (2018), inspired by neural Turing machines (Graves, Wayne, & Danihelka, 2014). Instead of using a pre-trained pairwise logistic regression model as local pairwise model in stage 1, like Laokulrat et al. (2015), Meng and Rumshisky (2018) employ a Long Short-Term Neural (LSTM) network classifier (Hochreiter & Schmidhuber, 1997). In the second stage, to classify the TLink relation for a candidate pair, they use the surrounding TLink predictions from the pre-trained model as input features in their final model. Another difference with Laokulrat et al. (2015) is that Meng and Rumshisky (2018) train their two stages jointly.

So for both Laokulrat et al. (2015) and Meng and Rumshisky (2018), the second phase uses no explicit knowledge about temporal reasoning (like reasoning rules). The label-label associations are learned from the data. An advantage of this is that the model can explicitly learn to correct mistakes of the local pairwise model from stage 1, which improves model performance. A disadvantage is that there are no guarantees on consistency.

5.3.4 Random Restart Hill Climbing (RRHC)

A less conventional TR approach was used by McClosky and Manning (2012), who addressed the task of temporal knowledge base population (KBP) slot filling (Ji, Grishman, Dang, Griffitt, & Ellis, 2010). Here the focus lies on finding the temporal bounds of a certain subset of events called *fluents*, which are semantic relations between entities that hold for a certain period of time, like *attends-school*, or *has-parent*. Their aim was to find the bounds of such events by detecting the following four types of TLinks (called *meta-relations* in the original paper) between the events and time expressions: BEGUN BY, ENDED BY, DURING (called START, END, and START AND END in the original paper), and a class UNRELATED. They employ local pairwise classifiers to obtain scores for each relation type, that are combined in a joint inference to obtain a globally consistent prediction using random restart hill climbing (RRHC). They define a global scoring function to score each set of predictions, which takes into account temporal consistency, and also the local scores from the pairwise classifiers. They find good scoring solution by iterating through all TE pairs, and at each pair adding the TLink that results in the setting with the highest score. Since this procedure depends on the initial order through which they randomly restart this procedure ten times, and pick the setting that gives the highest score from the ten final settings. This approach has not yet been evaluated in a more elaborate TIE setting, using a wider range of event types and temporal relation types.

5.3.5 Integer Linear Programming (ILP)

Another, more widely used, method to combine locally predicted TLink scores into a consistent temporal graph is integer linear programming (ILP). This technique was first exploited for TR by Bramsen et al. (2006b). They experiment with two greedy inference strategies: (1) following reading order, and (2) by in order of confidence, as described in the previous section. They also experimented with exact inference using integer linear programming (ILP). To formulate the problem as an integer linear program, it should be represented as a linear objective possibly extended with a set of linear constraints. Solving an ILP is in principle NP-complete, however there exist many efficient (often approximate) solvers (Berkelaar, Eikland, Notebaert, et al., 2004; Makhorin, 2008; Gu, Rothberg, & Bixby, 2012). Bramsen et al. (2006b) formulate the objective as the sum of all scores of the pairwise classifiers (for BEFORE, AFTER, and INCOMPARABLE). Additionally, they model three constraints:

1. Each segment pair is assigned only one label. (**mutual exclusivity**)
2. The BEFORE and AFTER relations follow transitivity. (**transitivity**)
3. Each segment is connected through at least one edge other than INCOMPARABLE. (**connectivity**)

Their experimental results show that using ILP performs better than greedy approaches like BF and NRO. Similar results were obtained as well by Mirroshandel and Ghassem-Sani (2012).

Chambers and Jurafsky (2008a) use a similar approach using TimeML-style data, predicting the same TLink types as Bramsen et al. (2006b), and also modeling transitivity

through ILP inference, but considering TLinks between events (EE-R) instead of between text segments.

They also show that using TR during prediction is more effective on densely connected temporal graphs. To obtain a densely connected temporal graph from the initially sparsely annotated annotations, they apply an extensive temporal closure to infer new EE-R. To compute the closure, they also exploit other TLink types (e.g., INCLUDES, and SIMULTANEOUS) and also the other relation categories: TE-R, DCT-R, and TT-R, later also used by others Tatu and Srikanth, Denis and Muller, Laokulrat et al., Ning et al. (2008, 2011, 2015, 2018a). The importance of densely connected graphs for TR-MI was also recognized by later research (Denis & Muller, 2011; Do, Lu, & Roth, 2012; Leeuwenberg & Moens, 2017).

Both Bramsen et al. (2006b), and Chambers and Jurafsky (2008a) only used two TimeML relations in their joint inference, working on a restricted problem setting, compared to using all twelve TLink types. When taking into account all twelve TLink types TR becomes much more computationally complex. This complexity problem is addressed by Denis and Muller (2011). In contrast to earlier approaches, they propose to formulate TR-based prediction using point-algebra instead of Allen algebra, exploiting the mapping proposed by Vilain et al. (1990), shown earlier in the lattice of Figure 6. This is done by translating the TLinks between intervals into relations between start and end points ($>$, $<$, and $=$). Their ILP objective maximizes the score from local pairwise classifiers, similar to Bramsen et al. (2006b), except that the ILP decision variables correspond to decisions about the translated point-wise relations, resulting in four times less variables. Because there are also fewer point-wise relation types (three) compared to interval relation types (twelve), they need a factor fifty fewer constraints to model the same reasoning fragment. After solving the ILP, the resulting point-wise decisions are translated back to TimeML interval relations. This shows that exploiting the connection between interval and point space greatly increases the computational efficiency of TR-based prediction with ILP. We argue that this method is very important for practical TIE systems. Compared to earlier ILP formulations, computational efficiency is gained, while at the same time expressiveness is increased. The full conjunctive sub-fragment discussed in section 3 is covered instead of just covering a handful of basic transitivity rules as is often done. Interestingly, more recently Kerr et al. (2014) used a quite large set of transitivity rules, using ILP to construct an ensemble from many local pairwise TLink extraction models, showing clear improvements. Although they used a large set of transitivity rules, they did not exploit the efficient point-algebraic formulation by Denis and Muller (2011).

A second modification to save computation has to do with temporal reasoning on subgroups of events, instead of all events in the document. The intuition is that people describe events in time frames, also called *narrative containers* (Pustejovsky & Stubbs, 2011), which has been adopted in later annotation works, and their corresponding systems (Styler IV et al., 2014; Bethard et al., 2015, 2016, 2017). This intuition of narrative containers is similar to the segments of Bramsen et al. (2006b), that correspond to larger phrases instead of individual events. Denis and Muller (2011) obtain these sub-graphs from ground-truth structures of connected events, and through events that correspond to the same temporal expression, and show how this can be used for more efficient TR.

5.3.6 Markov Logic Networks (MLN)

Another important method that has been proposed for TR-based prediction are Markov Logic Networks (MLN) (Richardson & Domingos, 2006). These were first explored for TIE by Yoshikawa et al. (2009), and later also by Ling and Weld (2010), and Ha et al. (2010). An major difference between MLN and the inference methods mentioned earlier, is that instead of combining locally trained models in a global inference setting, MLN also exploit the temporal constraints during training. Also, the weights for TR constraints can be learned, allowing the model to also learn soft correlations between TLinks, instead of hard rules.

To set up a MLN for a discriminative prediction problem, like predicting TLinks between events, one has to: (1) define a set of hidden first-order predicates that are observable during training that you want to predict (e.g., TLinks between different events), (2) define a set of observed predicates, available at both training and test-time (e.g., event features), and (3) define association rules among the predicates, which will get assigned a weight (e.g., feature-label associations, and label-label associations, like transitivity of certain TLinks). And (4), once the MLN is defined, a training and inference regime has to be determined to estimate probabilities for the hidden predicates and the association rules. This last step is often provided by MLN interpreters (Niu, Ré, Doan, & Shavlik, 2011). Yoshikawa et al. (2009) model different transitivity rules connecting the EE-R, TE-R, and DCT-R TLinks, and show that using MLN to incorporate TR outperforms local models that use no TR during prediction. MLN have been less popular in more recent works for scalability reasons (Leeuwenberg & Moens, 2017; Mojica & Ng, 2016). MLN constraints are soft (predicate assignments can become less likely, but not impossible), in contrast to approaches that model hard constraints, like ILP, that allow cutting off large areas from the search space to find solutions efficiently.

5.3.7 Structured Perceptron with Integer Linear Programming (SP+ILP)

Another approach to exploit TR in both training and prediction was proposed by Abend et al. (2015), in the domain of cooking recipes. They learn a global model formulating TIE as a structured learning problem. For this they combine an averaged structured perceptron (Freund & Schapire, 1999) with ILP inference. Abend et al. (2015) focused only on precedence relations between events, which was sufficient for cooking recipes. Similar approaches but for more extensive TimeML-based relations were proposed by Leeuwenberg and Moens (2017) in the clinical domain, and by Ning et al. (2017)[3] in the news domain. In these structured learning approaches, a scoring function is learned that scores groups of TLink assignments rather than single TLink assignments, as local models do. Model inference then corresponds to finding the assignment of TLinks with the highest overall score by the model. The naive inference method is to enumerate all possible TLinks assignments for a document, and pick the assignment with the highest score, which is usually highly computationally intractable. To formulate a more efficient inference procedure ILP is used

3. Which was later extended to deal with sparse annotations (Ning, Yu, Fan, & Roth, 2018d), jointly reason with causal relations (Ning et al., 2018a), and to include statistical knowledge from other resources (Ning, Wu, Peng, & Roth, 2018b).

to constrain the search space, similar to the approaches mentioned in the previous section. During training of the structured perceptron, the same ILP-style inference is used.

There are a few differences between the three SP+ILP approaches: Abend et al. (2015) focused on precedence relations between the events only. For this reason, their objective was to find a single chain of relations in which each event is visited only once, whereas the other two works use a more extensive rule set, making inference more complex. Also, Leeuwenberg and Moens (2017) besides hard-coded transitivity rules, also exploited soft learned label-label constraints. And, Ning et al. (2017) used an even more extensive transitivity table for TR, including more expressive rules that infer also some general Allen relations (disjunctions of TLinks), resulting in more expressive TR (class II). Similar to most approaches all works prune the total set of TLink candidate pairs to reduce computational complexity, and in all cases it was reported that TR during both training and prediction generally performs better compared to combining local classifiers with ILP.

5.3.8 Direct Timeline Models (DTLM)

Recently, a new type of approach to temporal event ordering was proposed by Leeuwenberg and Moens (2018a). Instead of predicting TLinks among events and temporal expressions, their model directly predicts the start and end points of events. An advantage of this approach is that it is fast, as predicting start and end points for each event is linear in the number of events, in contrast to predicting a set of TLinks, which is quadratic or requires pruning. To train their model they exploit TR to convert TLinks to sets of point-algebraic constraints. The loss function to train the model represents the distance that start and end points of events still need to shift to make all annotated temporal order relations valid on the predicted timeline. Another advantage of this approach is that its predicted timelines are consistent by definition. The main limitation of the approach is that there is no probabilistic interpretation of confidence for predictions, which is mentioned as future work.

5.4 Temporal Reasoning during Model Evaluation (TR-ME)

Initially, temporal information extraction systems were evaluated using either accuracy or F1-measure of extracted TLinks. Setzer et al. (2003) proposed to exploit TR to address a problem of straightforward calculation of accuracy or F1-measure:

- The same temporal situation can be represented using different TLinks (e.g., A before B represents the same situation as B after A with different labels).

To counter this problem they proposed to apply a temporal closure using all TLink types before evaluating F1-measure. However, this way, all TLinks are weighted equally, predicted and inferred TLinks. Tannier et al. (2008) addressed this problem by evaluating only with regard to *core relations*. From a set of TLinks, the core relations can be obtained by removing relations one-by-one, for as long as the inferable set of relations does not change (i.e., no information is lost). A problem with this approach however is that when comparing only core relations, not all inference information is captured, as already pointed out by Tannier et al. (2008) and Tannier and Muller (2011).

UzZaman and Allen (2011) have proposed a metric that deals with this issue, called temporal-awareness. They calculate a harmonic mean of precision and recall, i.e., an F-

score. A crucial difference with Setzer et al. (2003), and Tannier et al. (2008) is that to calculate their precision and recall they do not modify the original relations, but rather change the criterion on whether a relation is correctly classified with regard to the reference or not, using TR. Their precision metric is calculated as the percentage of system relations that can be *verified* from the reference relations using TR. Recall is calculated as the percentage of reference relations that can be verified from the system relations, using TR. To perform TR, they exploit TimeGraph (Miller & Schubert, 1990), an efficient TR algorithm based on the mapping between intervals and point algebra mentioned in Figure 6. TimeGraph conducts TR in the non-disjunctive sub-fragment of Allen's algebra, i.e., expressivity class III. The temporal-awareness is now used widely for evaluating TIE systems.

5.5 Overview of TR in TIE

There are some striking differences between the usage of TR in the different phases of model construction. During annotation, TR has shown positive results, and can help to reduce annotation work, and ensure consistency in the annotations. However, exploiting TR is not yet common practice in corpus construction.

In Table 2, we construct an overview of - to our knowledge - all TIE systems described in the literature that employ some form of interval or point-based reasoning for event ordering in the past three decades. We can see in the overview that in earlier approaches fewer TLink categories have been used for TR, focusing mostly on precedence relations (P) between event-event pairs (EE-R). In later approaches more relation types are predicted, like temporal inclusion (I), temporal equivalence (E), and overlap relations (O). This shows that the research focus in the community slowly grows in the direction of the challenge of full TIE, where *all* temporal cues from the text are extracted and combined at the same time, making TR an increasingly important aspect of TIE systems.

If we look at the use of TR to expand the training data (TR-DE) we observe mixed effects. Approaches that report high improvements above 10% improvement (Mani et al., 2006, 2007; Tatu & Srikanth, 2008) do so while splitting the training and test set on the relation instance level after TR. When splitting on the document-level, which is more realistic setting, the improvements are much smaller or even negative (Mani et al., 2007; Tatu & Srikanth, 2008; Nikfarjam et al., 2013; Mirza, 2014).

TR-based prediction approaches (TR-TP) are reported to outperform those that do not exploit TR, where ILP based approaches generally outperform greedy approaches (Bramsen et al., 2006b; Denis & Muller, 2011; Mirroshandel & Ghassem-Sani, 2012). A trend also seen in the table, is that more systems exploit TR not only during prediction (NRO, BF, ILP, SLI, RRHC), but also during training (SI, MLN, SP+ILP, DTLM), as this has shown to improve performance even further (Leeuwenberg & Moens, 2017; Ning et al., 2017), stressing the importance of integration of TR in TIE models.

The expressivity of the TR-based prediction approaches (in column TR-TP) mostly concerns basic Allen relations (class I). Ning et al. (2017, 2018a) extend this to transitivity rules with disjunctions of Allen relations in the conclusion (class II), but they still perform reasoning with the interval-level transitivity table. The vast majority of TR approaches that go beyond Allen's composition table of basic relations (from class I or II to class III) in TR expressiveness almost all perform reasoning in point algebra to remain tractable. This

Table 2: Overview of event ordering TIE systems using interval or point-based TR. The first column shows the reference, the second the types of the predicted temporal interval relations: precedence relations (P: *precedes*, *preceded by*, *meets*, *met by*) inclusion relations (I: *during*, *contains*, *starts*, *started by*, *ends*, *ended by*), overlap relations (O: *overlap*, *overlapped by*) or equivalence relations (E: *equals*). The next three columns indicate whether TR was used for data expansion (TR-DE), during training or prediction (TR-TP), or for model evaluation (TR-ME), where roman numerals indicate the expressivity class from section 3. The last column shows among what types of entities the relations are predicted (from section 4), where † indicates if ground truth relations were used in TR. If a reference directly evaluated the effect of TR, we report baseline score(s) and the change in score due to usage of TR (reported in %) for each experiment (separated by /). Scores and improvements are incomparable across references, as datasets, tasks, and evaluation metrics vary.

Reference	Rel. Types	TR-DE	TR-TP	TR-ME	Candidate Pairs
Mani et al. (2006)	P, I, E	✓$^{76+11}$	-	-	EE, TE-R
Mani et al. (2007)	P, I, E	✓$^{60+14/-11}$	-	-	EE, TE-R
Bramsen et al. (2006a)	P	✓	BF_I	-	SS-R
Bramsen et al. (2006b)	P	✓	NRO_I^{72+2}, BF_I^{+6}, ILP_I^{+12}	-	SS-R
Chambers and Jurafsky (2008a)	P	✓	ILP_I^{72+2}	-	EE, TE†, TT†-R
Verhagen and Pustejovsky (2008)	P, I, E	✓	$P\text{-}HOC_I$	-	EE-R
Tatu and Srikanth (2008)	P, I, E	✓$^{57+1/+15}$	$P\text{-}HOC_I^{50-3}$	-	EE, TE†, TT†-R
Yoshikawa et al. (2009)	P, I, E	-	MLN_I^{67+2}	-	EE, TE, DCT-R
Ling and Weld (2010)	P, O	-	MLN_I	-	TE-R
Ha et al. (2010)	P, O	-	MLN_I	-	EE, TE-R
Denis and Muller (2011)	P, I, E	✓	NRO_I^{11+14}, ILP_{III}^{+30}	-	EE, TE†, TT†-R
Mirroshandel and Ghassem-Sani (2012)	P, I, E	-	$BF_I^{48+3/49+1}$, $ILP_I^{+4/+2}$	-	EE-R
Do et al. (2012)	P, O	✓	ILP_I	-	EE, TE, DCT-R
McClosky and Manning (2012)	I	-	$RRHC_I^{71+1}$	-	TE-R
Costa and Branco (2013)	P, O	✓	NRO_I	-	EE, TE, DCT, TT-R
Nikfarjam et al. (2013)	P, O	✓$^{63+1}$	-	-	TE-R
Cheng et al. (2013)	P, O	✓	$P\text{-}HOC_I$	✓$_{III}$	EE,TE-R
Sun (2014)	P, O	-	$P\text{-}HOC_I$	✓$_{III}$	EE, TE-R
Chambers et al. (2014)	P, I, E	-	SLI_I^{32+8}	-	EE, TE, DCT, TT-R
Mirza (2014)	P, I, E	✓$^{48-2}$	-	✓$_{III}$	EE, TE, DCT-R
Kerr et al. (2014)	P, I, E, O	-	ILP_{III}	✓$_{III}$	EE, TE, DCT-R
Laokulrat et al. (2015)	P, I, E	✓	SI_I^{69+1}	✓$_{III}$	EE, TE, DCT, TT†-R
Abend et al. (2015)	P, E	-	$SP+ILP_I^{66+4/+5}$	✓$_{I}$	EE-R
Mirza and Tonelli (2016)	P, I, E	✓	$SLI_I^{61+1/49+2}$	✓$_{III}$	EE, TE, DCT-R
Li et al. (2016)	P, O	-	$ILP_I^{62+2/+6}$	-	EE-R
Cornegruta and Vlachos (2016)	I	-	-	✓$_{III}$	TE-R
Meng et al. (2017)	P, I, E, O	-	$P\text{-}HOC_I^{52+1/+4}$	✓$_{III}$	EE, TE, DCT, TT†-R
Leeuwenberg and Moens (2017)	P, I, O	✓	ILP_I^{83+1}, $SP+ILP_I^{+2}$	✓$_{III}$	EE, TE, DCT-R
Ning et al. (2017)	P, I, E	✓	ILP_{II}^{57+5}, $SP+ILP_{II}^{+10}$	✓$_{III}$	EE, TE, DCT-R
McDowell et al. (2017)	P, I, E	-	SLI_I^{49+2}	-	EE, TE, DCT-R
Ning et al. (2018a)	P, I, E	✓	$SP+ILP_{II}^{46+4/+5}$	✓$_{III}$	EE, TE, DCT, TT†-R
Meng and Rumshisky (2018)	P, I, E	-	SI_I^{53+1}	-	EE, TE, DCT, TT†-R
Leeuwenberg and Moens (2018a)	P, I, E	-	$DTLM_{III}$	✓$_{III}$	EE, TE, DCT-R

is observed in all areas of TIE: data expansion (Verhagen, 2005), training and prediction (Denis & Muller, 2011; Leeuwenberg & Moens, 2018a), and for evaluation (UzZaman & Allen, 2011). This indicates the importance of exploiting the point-interval mapping when considering practical systems, where expressivity and efficiency are both important.

It can be seen that during evaluation the expressivity of TR is frequently of class III. This is because the temporal awareness metric by (UzZaman & Allen, 2011) was adopted in the TempEval challenges (UzZaman et al., 2013; Minard et al., 2015; Bethard et al., 2016, 2017), and hence became a standard evaluation metric to use.

6. Future Directions and Discussion

In this section we discuss the results from the survey and aim to point out areas that have been relatively unexplored or that we believe are promising for future work.

In the previous section, we observed that TR during prediction improves over using no TR, global methods outperform local greedy methods, and integrating TR in both prediction and training can improve model performance even further.

Efficiency of TR has not been compared explicitly across models in the existing TIE literature. However, it is used as motivation for many works to choose for a certain method. In general, greedy methods and sieve-based methods (BF, NRO, SI), which look for local optimal solutions, are faster than global methods like ILP, MLN, and RRHC, possibly at the cost of performance. There has been some work on comparing efficiency of ILP-style inference and MLN, which suggests that ILP is more efficient for currently existing solvers (Mojica & Ng, 2016). Also, SP+ILP methods are generally slower in training time than ILP-based methods, as their more complex inference procedure is also performed during training, however as seen in Table 2, it also often provides further improvements. To choose the degree of TR may depend on your dataset as well, as predicting densely connected graphs generally benefits more from TR. As mentioned in section 3, irrespective of the degree of TR, point-based TR methods are generally faster than equally expressive interval-based TR methods.

In this light, we observe that more recent works focus more on annotating and predicting relations between start and end points of events, rather than relations between events or intervals (Reimers et al., 2016, 2018; Ning et al., 2018c; Ning, Zhou, Feng, Peng, & Roth, 2018e; Leeuwenberg & Moens, 2018a). We believe this is a promising and important change in perspective also with regard to TR. To increase expressivity of a TR component up to class III while remaining tractable reasoning in point-algebra is crucial. Additionally, representing the temporality of events by their start and end points provides flexibility when combining different annotation schemes, as most schemes can be converted to point-algebra.

Also, this flexibility could be very beneficial when incorporating other types of reasoning. Many questions that involve temporality cannot be solved using TR alone, but require other types of semantic reasoning about events and entities (Höffner et al., 2017; Pampari et al., 2018; Suster & Daelemans, 2018), including (but not limited to) co-reference (Do et al., 2012), spatial reasoning, which has strong similarities with temporal reasoning (Guesgen, 1989; Mukerjee & Joe, 1990; Freksa, 1992b; Walsh, 2003), and causal reasoning and extraction (Bethard, Corvey, Klingenstein, & Martin, 2008; Mirza & Tonelli, 2014; Mirza, 2014;

Mirza & Tonelli, 2016; Mostafazadeh, Grealish, Chambers, Allen, & Vanderwende, 2016; Dunietz, Levin, & Carbonell, 2017; Ning et al., 2018a).

To incorporate these different types of information into single TIE models we believe neural networks could be a suitable model class, as they are very flexible in combining multiple tasks, and incorporating background knowledge like typical event orders (Chklovski & Pantel, 2004; Pichotta & Mooney, 2016), durations (Vempala, Blanco, & Palmer, 2018), or times of the day (Noro, Inui, Takamura, & Okumura, 2006). Currently, many neural TIE approaches follow the pairwise TLink classification paradigm, and are based on LSTM (Tourille, Ferret, Neveol, & Tannier, 2017; Cheng & Miyao, 2017; Leeuwenberg & Moens, 2018b; Meng et al., 2017; Choubey & Huang, 2017; Lin, Miller, Dligach, Amiri, Bethard, & Savova, 2018; Lin, Miller, Dligach, Bethard, & Savova, 2019), convolutional neural networks (CNN) (Dligach, Miller, Lin, Bethard, & Savova, 2017; Lin, Miller, Dligach, Bethard, & Savova, 2017), tree-based LSTM networks (Galvan, Okazaki, Matsuda, & Inui, 2018), and attention networks (Liu, Wang, Chaudhary, & Liu, 2019). However, currently the exploitation of TR for neural TIE has been very limited in neural TIE systems leaving room for future research.

Another area we would like to address is the challenging area of cross-document TIE. For TR in cross-document TIE temporal cues from multiple documents need to be combined, stressing the importance of computational efficiency. Cross-document TIE has not been discussed elaborately in this survey as the amount of TR in this research area has been very limited, possibly for this very reason of computational efficiency. Barzilay and McKeown (2005) were one of the first to dive into this area of research, and used *iconicity*, the heuristic that in narrative texts events are often mentioned in chronological order, to temporally order sentences in a multi-document summarization task. There has been work on generating course grained entity-focused timelines from multiple news articles as a means to do multi-document summarization (Yan, Kong, Huang, Wan, Li, & Zhang, 2011; Zhao, Guo, Yan, He, & Li, 2013; Lin, Efron, Wang, & Sherman, 2014; Wang, Shou, Chen, Chen, & Mehrotra, 2015; Althoff, Dong, Murphy, Alai, Dang, & Zhang, 2015). However, the focus of most of the research in this area lies mainly on the informativeness of sentences for the summary, and less on the temporal aspect. In most cases, TIE is very limited, and it is simply assumed that all events mentioned in each text occur at the document-creation time, leaving many opportunities for TIE and TR. Besides computational efficiency, a challenge in a cross-document TIE setting is that event mentions of the same event should be linked across documents, called cross-document event co-reference. Event co-reference is strongly connected to TIE as a single event can only occur at a single time. Do et al. (2012) modeled this principle connecting TIE and event co-reference in their TR-based prediction using ILP. However, this was in a within-document setting. This could be a good starting point for cross-document ILP formulations. Minard et al. (2015) provided a setup for evaluating this problem in a shared task. However, TR has not yet been explored by the participants of this shared task. Also the ECB+ corpus (Cybulska & Vossen, 2014), which contains event co-reference and temporal information annotations across documents could be an interesting resource.

With regard to general TIE, considering both cross-document but also within-document relations, we can observe from the overview in the previous section that the state-of-the-art systems using TR focus mainly on *ordering* events, and *anchoring* events to temporal

expressions, using TimeML-style data. Since TimeML only annotates duration cues that are definite, quantified and explicit it does not contain implicit or background event duration information, nor explicit indefinite quantification (*long meeting*). There have been TIE systems that also use implicit duration information (Pan et al., 2006b; Reimers et al., 2018). However, these systems do not exploit TR, nor TimeML-style annotations, leaving a gap in the literature: systems that are able to combine different data sources, like TimeML and event duration annotations (Pan et al., 2011; Reimers et al., 2016). Combining different types of information and learning from different data sources presents new challenges for TR. For the TR aspect it might be interesting to explore temporal constraint networks, that are able to deal with quantification (Dechter & Cohen, 2003). Two systems for which we believe further integration of event duration information would be straightforward are: (1) the direct timeline models by Leeuwenberg and Moens (2018a), and (2) the work of Reimers et al. (2018), as both approaches already predict and combine event position and duration.

7. Conclusions

We presented a comprehensive survey on how temporal reasoning mechanisms can be exploited for temporal information extraction, covering the literature of the last three decades in this research area.

To explain the complexity of the temporal information that is present in language we provided an exemplified overview of the different types of temporal information present in language: absolute v.s. relative cues, definite and indefinite cues, implicit cues, and background knowledge. Many types of temporal cues, like explicit event position cues, and event durations can already be extracted by different types of temporal information extraction systems from the literature.

There appears to be a trend towards more complete TR, going from only reasoning about EE-R, towards including also TE-R, DCT-R, and even TT-R. Although still no systems seem to combine all temporal cues as of yet. For example, information about the relative ordering of events and quantified duration information have not yet been combined in TR, although data is available, accommodating room for future work on joint models.

To provide a back-bone for the state-of-the-art TIE systems using TR, a comprehensive explanation of the most widely used temporal reasoning frameworks in temporal information extraction systems was given. We reviewed Allen's interval algebra and its relation to point algebra in detail, giving insight into the considerations with regard to expressiveness of temporal reasoning and computational efficiency. We argue that for obtaining practical systems point-algebraic approaches for TR are preferred, as they strike a good balance between expressiveness and efficiency instead of reasoning directly with Allen's interval relations.

In the core of the survey, we reviewed the different methods to exploit temporal reasoning for constructing temporal information extraction models and distilled the most widely confirmed conclusions. TR during annotation has been proposed already early on, and is effective for ensuring connectivity and consistency in annotations but has been used to a limited degree in existing corpora. To expand the often only sparsely annotated TimeML data, a transitive closure is used frequently to densify the annotated temporal graphs. This

has been found mostly beneficial when TR is also used during model inference, as TR during model inference appears to work better on densely connected temporal graphs. Generally, usage of TR for model inference appears to increase model performance. Some approaches use TR only during training, and some both during training and inference. Inference-only approaches include: BF, ILP, RRHC, and SLI, and approaches that use TR also during training include: MLN, SI, SP+ILP, and DTLM. The last category of approaches has been researched most recently and has been reported to perform better than inference-only approaches. For evaluation the research community has converged on using the TR-based temporal awareness measure.

In closing, it is clear that TR is crucial for TIE, and widely used in all aspects of model construction. However, most current research on TIE still addresses sub-fragments of the complete TIE problem, focusing on extraction of specific types of temporal cues, instead of extracting all cues jointly which would allow them to complement each other. Consequently, it remains an open research question how to perform efficient and expressive TR involving all types of temporal cues. We believe to answer this question, a flexible, expressive and efficient reasoning framework is required. For this, we believe important directions of research are point-based reasoning approaches, striking a good balance between efficiency and expressiveness, and deep learning methods, that facilitate flexibility in model construction, multi-task learning, and sharing of representations.

Acknowledgements

The authors thank Geert Heyman and the reviewers for their constructive comments which helped us to improve the paper. This work was supported by the ERC Advanced Grant CALCULUS (788506), the KU Leuven project MARS (C22/15/16), and by the IWT-SBO project ACCUMULATE (150056).

References

Abend, O., Cohen, S. B., & Steedman, M. (2015). Lexical event ordering with an edge-factored model. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1161–1171. ACL.

Allen, J. F. (1983). Maintaining knowledge about temporal intervals. In *Communications of the ACM*, pp. 832–843. ACM.

Althoff, T., Dong, X. L., Murphy, K., Alai, S., Dang, V., & Zhang, W. (2015). TimeMachine: Timeline generation for knowledge-base entities. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 19–28. ACM.

Augusto, J. C. (2005). Temporal reasoning for decision support in medicine. In *Artificial Intelligence in Medicine*, Vol. 33, pp. 1–24. Elsevier.

Barzilay, R., & McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. In *Computational Linguistics*, Vol. 31, pp. 297–328. MIT Press.

Berkelaar, M., Eikland, K., Notebaert, P., et al. (2004). lpsolve: Open source (mixed-integer) linear programming system. In *Eindhoven University of Technology*.

Bethard, S., Corvey, W. J., Klingenstein, S., & Martin, J. H. (2008). Building a corpus of temporal-causal structure. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 908–915. ELRA.

Bethard, S., Derczynski, L., Savova, G., Pustejovsky, J., & Verhagen, M. (2015). Semeval-2015 Task 6: Clinical TempEval. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 806–814. ACL.

Bethard, S., & Parker, J. (2016). A semantically compositional annotation scheme for time normalization. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 3779–3786, Paris, France. ELRA.

Bethard, S., Savova, G., Chen, W.-T., Derczynski, L., Pustejovsky, J., & Verhagen, M. (2016). Semeval-2016 Task 12: Clinical TempEval. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 1052–1062. ACL.

Bethard, S., Savova, G., Palmer, M., & Pustejovsky, J. (2017). SemEval-2017 Task 12: Clinical TempEval. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 565–572, Vancouver, Canada. ACL.

Boguraev, B., & Ando, R. K. (2007). Effective use of TimeBank for TimeML analysis. In *Annotating, Extracting and Reasoning about Time and Events*, pp. 41–58. Springer.

Bramsen, P., Deshpande, P., Lee, Y. K., & Barzilay, R. (2006a). Finding temporal order in discharge summaries. In *Proceedings of the AMIA Annual Symposium*, Vol. 2006, p. 81:5. American Medical Informatics Association.

Bramsen, P., Deshpande, P., Lee, Y. K., & Barzilay, R. (2006b). Inducing temporal graphs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 189–198. ACL.

Campos, R., Dias, G., Jorge, A. M., & Jatowt, A. (2015). Survey of temporal information retrieval and related applications. *ACM Computing Surveys*, *47*(2), 15.

Cassidy, T., McDowell, B., Chambers, N., & Bethard, S. (2014). An annotation framework for dense event ordering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Vol. 2, pp. 501–506. ACL.

Chambers, N., Cassidy, T., McDowell, B., & Bethard, S. (2014). Dense event ordering with a multi-pass architecture. In *Transactions of the Association for Computational Linguistics*, Vol. 2, pp. 273–284.

Chambers, N., & Jurafsky, D. (2008a). Jointly combining implicit constraints improves temporal ordering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 698–706. ACL.

Chambers, N., & Jurafsky, D. (2008b). Unsupervised learning of narrative event chains. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 789–797.

Chambers, N., Wang, S., & Jurafsky, D. (2007). Classifying temporal relations between events. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 173–176. ACL.

Chang, A. X., & Manning, C. D. (2012). SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Vol. 2012, pp. 3735–3740. ELRA.

Cheng, F., & Miyao, Y. (2017). Classifying temporal relations by bidirectional LSTM over dependency paths. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Vol. 2, pp. 1–6. ACL.

Cheng, Y., Anick, P., Hong, P., & Xue, N. (2013). Temporal relation discovery between events and temporal expressions identified in clinical narrative. In *Journal of Biomedical Informatics*, Vol. 46, pp. S48–S53. Elsevier.

Chklovski, T., & Pantel, P. (2004). VerbOcean: Mining the Web for fine-grained semantic verb relations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 33–40. ACL.

Choi, E., Bahadori, M. T., Sun, J., Kulas, J., Schuetz, A., & Stewart, W. (2016a). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pp. 3504–3512.

Choi, E., Schuetz, A., Stewart, W. F., & Sun, J. (2016b). Using recurrent neural network models for early detection of heart failure onset. In *Journal of the American Medical Informatics Association*, Vol. 24, pp. 361–370. Oxford University Press.

Choubey, P. K., & Huang, R. (2017). A sequential model for classifying temporal relations between intra-sentence events. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1796–1802. ACL.

Cornegruta, S., & Vlachos, A. (2016). Timeline extraction using distant supervision and joint inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1936–1942. ACL.

Costa, F., & Branco, A. (2013). Temporal relation classification based on temporal reasoning. In *Proceedings of IWCS*, pp. 59–70.

Cybulska, A., & Vossen, P. (2014). Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 4545–4552. ELRA.

Dechter, R., & Cohen, D. (2003). Constraint processing.. Morgan Kaufmann.

Dechter, R., Meiri, I., & Pearl, J. (1991). Temporal constraint networks. In *Artificial Intelligence*, Vol. 49, pp. 61–95. Elsevier.

Denis, P., & Muller, P. (2011). Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1788–1793. AAAI Press.

Derczynski, L. (2016). Representation and learning of temporal relations. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 1937–1948. ACL.

Derczynski, L., Llorens, H., & Saquete, E. (2012). Massively increasing TIMEX3 resources: A transduction approach. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 3754–3761. ELRA.

Derczynski, L., Llorens, H., & UzZaman, N. (2013). TimeML-strict: clarifying temporal annotation. In *arXiv preprint arXiv:1304.7289*.

Derczynski, L., Strötgen, J., Maynard, D., Greenwood, M. A., & Jung, M. (2016). GATE-Time: Extraction of temporal expressions and event. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 3702–3708. ELRA.

Derczynski, L. R. (2017). *Automatically Ordering Events and Times in Text*. Springer.

Dligach, D., Miller, T., Lin, C., Bethard, S., & Savova, G. (2017). Neural temporal relation extraction. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Vol. 2, pp. 746–751.

Do, Q. X., Lu, W., & Roth, D. (2012). Joint inference for event timeline construction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 677–687. ACL.

Dunietz, J., Levin, L., & Carbonell, J. (2017). The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the Linguistic Annotation Workshop*, pp. 95–104.

Filannino, M., Brown, G., & Nenadic, G. (2013). ManTIME: Temporal expression identification and normalization in the TempEval-3 challenge. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, Vol. 2, pp. 53–57. ACL.

Fisher, M. D., Gabbay, D. M., & Vila, L. (2005). Handbook of temporal reasoning in artificial intelligence.. Vol. 1. Elsevier.

Freksa, C. (1992a). Temporal reasoning based on semi-intervals. In *Artificial Intelligence*, Vol. 54, pp. 199–227. Elsevier.

Freksa, C. (1992b). Using orientation information for qualitative spatial reasoning. In *Theories and Methods of Spatio-temporal Reasoning in Geographic Space*, pp. 162–178. Springer.

Freund, Y., & Schapire, R. E. (1999). Large margin classification using the perceptron algorithm. In *Machine Learning*, Vol. 37, pp. 277–296. Springer.

Galvan, D., Okazaki, N., Matsuda, K., & Inui, K. (2018). Investigating the challenges of temporal relation extraction from clinical text. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pp. 55–64.

Gennari, R., & Vittorini, P. (2016). Qualitative temporal reasoning can improve on temporal annotation quality: How and why. In *Applied Artificial Intelligence*, Vol. 30, pp. 690–719. Taylor & Francis.

Gennari, R., & Vittorini, P. (2017). Time out of joint in temporal annotations of texts: Challenges for artificial intelligence and human computer interaction. In *Proceedings of the AI*IA Workshop on Deep Understanding and Reasoning*, pp. 50–55. CEUR-WS. org.

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural Turing machines. In *arXiv preprint arXiv:1410.5401*.

Gu, Z., Rothberg, E., & Bixby, R. (2012). Gurobi optimizer reference manual, version 5.0. In *Gurobi Optimization Inc., Houston, USA*.

Guesgen, H. W. (1989). *Spatial Reasoning Based on Allen's Temporal Logic*. International Computer Science Institute Berkeley.

Ha, E. Y., Baikadi, A., Licata, C., & Lester, J. C. (2010). NCSU: Modeling temporal relations with Markov logic and lexical ontology. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 341–344. ACL.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. In *Neural Computation*, Vol. 9, pp. 1735–1780. MIT Press.

Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J., & Ngonga Ngomo, A.-C. (2017). Survey on challenges of question answering in the semantic web. *Semantic Web*, *8*(6), 895–920.

Ji, H., Grishman, R., Dang, H. T., Griffitt, K., & Ellis, J. (2010). Overview of the TAC 2010 knowledge base population track. In *Proceedings of the Third Text Analysis Conference*, Vol. 3, pp. 3–3.

Jung, H., Allen, J., Blaylock, N., De Beaumont, W., Galescu, L., & Swift, M. (2011). Building timelines from narrative clinical records: initial results based-on deep natural language understanding. In *Proceedings of the Workshop on Biomedical Natural Language Processing*, pp. 146–154. ACL.

Kerr, C., Hoare, T., Carroll, P., & Marecek, J. (2014). Integer-programming ensemble of temporal-relations classifiers. In *arXiv preprint arXiv:1412.1866*.

Kolomiyets, O., & Moens, M.-F. (2010). KUL: recognition and normalization of temporal expressions. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 325–328. ACL.

Kozareva, Z., & Hovy, E. (2011). Learning temporal information for states and events. In *Proceedings of the International Conference on Semantic Computing*, pp. 424–429. IEEE.

Krokhin, A., Jeavons, P., & Jonsson, P. (2003). Reasoning about temporal relations: The tractable subalgebras of Allen's interval algebra. In *Journal of the ACM*, Vol. 50, pp. 591–640. ACM.

Laokulrat, N., Miwa, M., & Tsuruoka, Y. (2014). Exploiting timegraphs in temporal relation classification. In *Proceedings of the Workshop on Graph-based Methods for Natural Language Processing*, pp. 6–14.

Laokulrat, N., Miwa, M., & Tsuruoka, Y. (2015). Stacking approach to temporal relation classification with temporal inference. In *Journal of Natural Language Processing*, Vol. 22, pp. 171–196.

Laparra, E., Xu, D., & Bethard, S. (2018a). From characters to time intervals: New paradigms for evaluation and neural parsing of time normalizations. In *Transactions of the Association of Computational Linguistics*, Vol. 6, pp. 343–356.

Laparra, E., Xu, D., Elsayed, A., Bethard, S., & Palmer, M. (2018b). SemEval 2018 Task 6: Parsing time normalizations. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 88–96. ACL.

Lee, K., Artzi, Y., Dodge, J., & Zettlemoyer, L. (2014). Context-dependent semantic parsing for time expressions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Vol. 1, pp. 1437–1447. ACL.

Leeuwenberg, A., & Moens, M.-F. (2017). Structured learning for temporal relation extraction from clinical records. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 1150–1158. ACL.

Leeuwenberg, A., & Moens, M.-F. (2018a). Temporal information extraction by predicting relative time-lines. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1237–1246, Brussels, Belgium. ACL.

Leeuwenberg, A., & Moens, M.-F. (2018b). Word-level loss extensions for neural temporal relation classification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 3436–3447. ACL.

Li, P., Zhu, Q., Zhou, G., & Wang, H. (2016). Global inference to Chinese temporal relation extraction. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 1451–1460. ACL.

Ligozat, G. (1996). A new proof of tractability for ORD-Horn relations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 395–401.

Lin, C., Miller, T., Dligach, D., Amiri, H., Bethard, S., & Savova, G. (2018). Self-training improves recurrent neural networks performance for temporal relation extraction. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pp. 165–176.

Lin, C., Miller, T., Dligach, D., Bethard, S., & Savova, G. (2017). Representations of time expressions for temporal relation extraction with convolutional neural networks. In *Proceedings of the Workshop on Biomedical Natural Language Processing*, pp. 322–327.

Lin, C., Miller, T., Dligach, D., Bethard, S., & Savova, G. (2019). A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction. In *Proceedings of the Clinical Natural Language Processing Workshop*, pp. 65–71.

Lin, J., Efron, M., Wang, Y., & Sherman, G. (2014). Overview of the TREC-2014 microblog track. Tech. rep., Maryland University College Park.

Lin, Y.-K., Chen, H., & Brown, R. A. (2013). MedTime: A temporal information extraction system for clinical narratives. In *Journal of Biomedical Informatics*, Vol. 46, pp. S20–S28. Elsevier.

Ling, X., & Weld, D. S. (2010). Temporal information extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 10, pp. 1385–1390.

Liu, S., Wang, L., Chaudhary, V., & Liu, H. (2019). Attention neural model for temporal relation extraction. In *Proceedings of the Clinical Natural Language Processing Workshop*, pp. 134–139, Minneapolis, Minnesota, USA. ACL.

Llorens, H., Chambers, N., UzZaman, N., Mostafazadeh, N., Allen, J., & Pustejovsky, J. (2015). SemEval-2015 task 5: QA TempEval - evaluating temporal information understanding with question answering. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 792–800. ACL.

Llorens, H., Derczynski, L., Gaizauskas, R. J., & Saquete, E. (2012). TIMEN: An open temporal expression normalisation resource. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 3044–3051. ELRA.

Makhorin, A. (2008). GLPK (GNU linear programming kit). In *http://www.gnu.org/s/glpk/glpk.html*.

Mani, I., Verhagen, M., Wellner, B., Lee, C. M., & Pustejovsky, J. (2006). Machine learning of temporal relations. In *Proceedings of the International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pp. 753–760. ACL.

Mani, I., Wellner, B., Verhagen, M., & Pustejovsky, J. (2007). Three approaches to learning TLINKS in TimeML. In *Technical Report CS-07–268, Computer Science Department*.

Mani, I., & Wilson, G. (2000). Robust temporal processing of news. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 69–76. ACL.

McClosky, D., & Manning, C. D. (2012). Learning constraints for consistent timeline extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 873–882. ACL.

McDowell, B., Chambers, N., Ororbia II, A., & Reitter, D. (2017). Event ordering with a generalized model for sieve prediction ranking. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, Vol. 1, pp. 843–853. ACL.

Meiri, I. (1996). Combining qualitative and quantitative constraints in temporal reasoning. In *Artificial Intelligence*, Vol. 87, pp. 343–385. Elsevier.

Meng, Y., & Rumshisky, A. (2018). Context-aware neural model for temporal information extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Vol. 1, pp. 527–536. ACL.

Meng, Y., Rumshisky, A., & Romanov, A. (2017). Temporal information extraction for question answering using syntactic dependencies in an LSTM-based architecture. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 887–896. ACL.

Miller, S., & Schubert, L. (1990). Time revisited. In *Computational Intelligence*, Vol. 6, pp. 108–118. Blackwell.

Miller, T., Bethard, S., Dligach, D., Lin, C., & Savova, G. (2015). Extracting time expressions from clinical text. In *Proceedings of the Workshop on Biomedical Natural Language Processing*, pp. 81–91, Beijing, China. ACL.

Minard, A.-L., Speranza, M., Agirre, E., Aldabe, I., van Erp, M., Magnini, B., Rigau, G., & Urizar, R. (2015). Semeval-2015 Task 4: Timeline: Cross-document event ordering. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 778–786. ACL.

Mirroshandel, S. A., & Ghassem-Sani, G. (2012). Towards unsupervised learning of temporal relations between events. In *Journal of Artificial Intelligence Research*, Vol. 45, pp. 125–163.

Mirza, P. (2014). Extracting temporal and causal relations between events. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics Student Research Workshop*, pp. 10–17.

Mirza, P. (2015). Recognizing and normalizing temporal expressions in Indonesian texts. In *Conference of the Pacific Association for Computational Linguistics*, pp. 135–147. Springer.

Mirza, P., & Tonelli, S. (2014). An analysis of causality between events and its relation to temporal information. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 2097–2106. ACL.

Mirza, P., & Tonelli, S. (2016). Catena: Causal and temporal relation extraction from natural language texts. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 64–75. ACL.

Mojica, L. G., & Ng, V. (2016). Markov Logic Networks for text mining: A qualitative and empirical comparison with integer linear programming. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 4388–4395. ELRA.

Mostafazadeh, N., Grealish, A., Chambers, N., Allen, J., & Vanderwende, L. (2016). CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Workshop on Events*, pp. 51–61.

Mukerjee, A., & Joe, G. (1990). *A Qualitative Model for Space*. Texas A and M University. Computer Science Department.

Nebel, B., & Bürckert, H.-J. (1995). Reasoning about temporal relations: a maximal tractable subclass of Allen's interval algebra. In *Journal of the ACM*, Vol. 42, pp. 43–66. ACM.

Negri, M., & Marseglia, L. (2004). Recognition and normalization of time expressions: ITC-irst at TERN 2004. In *Rapport interne, ITC-irst, Trento*.

Ng, J.-P., Chen, Y., Kan, M.-Y., & Li, Z. (2014). Exploiting timelines to enhance multi-document summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Vol. 1, pp. 923–933. ACL.

Nikfarjam, A., Emadzadeh, E., & Gonzalez, G. (2013). Towards generating a patient's timeline: extracting temporal relationships from clinical notes. In *Journal of Biomedical Informatics*, Vol. 46, pp. S40–S47. Elsevier.

Ning, Q., Feng, Z., & Roth, D. (2017). A structured learning approach to temporal relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1027–1037. ACL.

Ning, Q., Feng, Z., Wu, H., & Roth, D. (2018a). Joint reasoning for temporal and causal relations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2278–2288. ACL.

Ning, Q., Wu, H., Peng, H., & Roth, D. (2018b). Improving temporal relation extraction with a globally acquired statistical resource. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 841–851, New Orleans, Louisiana. ACL.

Ning, Q., Wu, H., & Roth, D. (2018c). A multi-axis annotation scheme for event temporal relations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1318–1328, Melbourne, Australia. ACL.

Ning, Q., Yu, Z., Fan, C., & Roth, D. (2018d). Exploiting partially annotated data for temporal relation extraction. In *Proceedings of Joint Conference on Lexical and Computational Semantics (*SEM)*, pp. 148–153. ACL.

Ning, Q., Zhou, B., Feng, Z., Peng, H., & Roth, D. (2018e). CogCompTime: A tool for understanding time in natural language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP)*, pp. 72–77, Brussels, Belgium. ACL.

Niu, F., Ré, C., Doan, A., & Shavlik, J. (2011). Tuffy: Scaling up statistical inference in Markov logic networks using an RDBMS. In *Proceedings of the VLDB Endowment*, Vol. 4, pp. 373–384. VLDB Endowment.

Noro, T., Inui, T., Takamura, H., & Okumura, M. (2006). Time period identification of events in text. In *Proceedings of the International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pp. 1153–1160. ACL.

Olex, A., Maffey, L., Morgan, N., & McInnes, B. (2018). Chrono at SemEval-2018 Task 6: A system for normalizing temporal expressions. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 97–101. ACL.

Pampari, A., Raghavan, P., Liang, J., & Peng, J. (2018). emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2357–2368. ACL.

Pan, F., Mulkar-Mehta, R., & Hobbs, J. R. (2006a). An annotated corpus of typical durations of events. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 77–82. ELRA.

Pan, F., Mulkar-Mehta, R., & Hobbs, J. R. (2006b). Learning event durations from event descriptions. In *Proceedings of the International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pp. 393–400. ACL.

Pan, F., Mulkar-Mehta, R., & Hobbs, J. R. (2011). Annotating and learning event durations in text. In *Computational Linguistics*, Vol. 37, pp. 727–752. MIT Press.

Pichotta, K., & Mooney, R. J. (2016). Learning statistical scripts with LSTM recurrent neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2800–2806.

Pustejovsky, J., Castano, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., & Radev, D. R. (2003a). TimeML: Robust specification of event and temporal expressions in text. In *New directions in Question Answering*, Vol. 3, pp. 28–34.

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al. (2003b). The TimeBank corpus. In *Corpus Linguistics*, Vol. 2003, p. 40. Lancaster, UK.

Pustejovsky, J., Lee, K., Bunt, H., & Romary, L. (2010). ISO-TimeML: An international standard for semantic annotation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Vol. 10, pp. 394–397.

Pustejovsky, J., & Stubbs, A. (2011). Increasing informativeness in temporal annotation. In *Proceedings of the Linguistic Annotation Workshop*, pp. 152–160. ACL.

Raghavan, P., Chen, J. L., Fosler-Lussier, E., & Lai, A. M. (2014). How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?. In *Proceedings of the AMIA Summits on Translational Science*, Vol. 2014, pp. 218–223. American Medical Informatics Association.

Real, L., Rademaker, A., Chalub, F., & de Paiva, V. (2018). Towards temporal reasoning in Portuguese. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. ELRA.

Reimers, N., Dehghani, N., & Gurevych, I. (2016). Temporal anchoring of events for the TimeBank corpus. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Vol. 1, pp. 2195–2204.

Reimers, N., Dehghani, N., & Gurevych, I. (2018). Event time extraction with a decision tree of neural classifiers. In *Transactions of the Association for Computational Linguistics*, Vol. 6, pp. 77–89. ACL.

Richardson, M., & Domingos, P. (2006). Markov logic networks. In *Machine Learning*, Vol. 62, pp. 107–136. Springer.

Schockaert, S., & De Cock, M. (2008). Temporal reasoning about fuzzy intervals. In *Artificial Intelligence*, Vol. 172, pp. 1158–1193. Elsevier.

Settles, B. (2012). Active learning. In *Synthesis Lectures on Artificial Intelligence and Machine Learning*, Vol. 6, pp. 1–114. Morgan & Claypool Publishers.

Setzer, A., Gaizauskas, R., & Hepple, M. (2003). Using semantic inferences for temporal annotation comparison. In *Proceedings of the International Workshop on Inference in Computational Semantics*.

Setzer, A., Gaizauskas, R., & Hepple, M. (2005). The role of inference in the temporal annotation and analysis of text. In *Language Resources and Evaluation*, Vol. 39, pp. 243–265. Springer.

Strötgen, J., & Gertz, M. (2010). HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 321–324. ACL.

Strötgen, J., & Gertz, M. (2015). A baseline temporal tagger for all languages. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 541–547. ACL.

Strötgen, J., & Gertz, M. (2016). Domain-sensitive temporal tagging. In *Synthesis Lectures on Human Language Technologies*, Vol. 9, pp. 1–151. Morgan & Claypool Publishers.

Styler IV, W. F., Bethard, S., Finan, S., Palmer, M., Pradhan, S., de Groen, P. C., Erickson, B., Miller, T., Lin, C., Savova, G., et al. (2014). Temporal annotation in the clinical domain. In *Transactions of the Association for Computational Linguistics*, Vol. 2, pp. 143–154. MIT Press.

Sun, W. (2014). Time will tell: Temporal reasoning in clinical narratives and beyond.. State University of New York at Albany.

Sun, W., Rumshisky, A., & Uzuner, O. (2013). Annotating temporal information in clinical narratives. In *Journal of Biomedical Informatics*, Vol. 46, pp. S5–S12. Elsevier.

Sun, W., Rumshisky, A., & Uzuner, O. (2015). Normalization of relative and incomplete temporal expressions in clinical narratives. In *Journal of the American Medical Informatics Association*, Vol. 22, pp. 1001–1008. Oxford University Press.

Sun, Y., Cheng, G., & Qu, Y. (2018). Reading comprehension with graph-based temporal-casual reasoning. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 806–817. ACL.

Suster, S., & Daelemans, W. (2018). CliCR: a dataset of clinical case reports for machine reading comprehension. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 1551–1563. ACL.

Tannier, X., & Muller, P. (2011). Evaluating temporal graphs built from texts via transitive reduction. In *Journal of Artificial Intelligence Research*, Vol. 40, pp. 375–413.

Tannier, X., Muller, P., et al. (2008). Evaluation metrics for automatic temporal annotation of texts. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 150–155. ELRA.

Tatu, M., & Srikanth, M. (2008). Experiments with reasoning for temporal relations between events. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 857–864. ACL.

Tissot, H., Roberts, A., Derczynski, L., Gorrell, G., & Del Fabro, M. D. (2015). Analysis of temporal expressions annotated in clinical notes. In *Proceedings of the Workshop on Interoperable Semantic Annotation*. ACL.

Tourille, J., Ferret, O., Neveol, A., & Tannier, X. (2017). Neural architecture for temporal relation extraction: A Bi-LSTM approach for detecting narrative containers. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Vol. 2, pp. 224–230.

UzZaman, N., & Allen, J. F. (2011). Temporal evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, HLT '11, pp. 351–356, Stroudsburg, PA, USA. ACL.

UzZaman, N., Llorens, H., Derczynski, L., Allen, J., Verhagen, M., & Pustejovsky, J. (2013). Semeval-2013 Task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Proceedings of the Joint Conference on Lexical and Computational Semantics and the International Workshop on Semantic Evaluation (*SEM-SemEval)*, Vol. 2, pp. 1–9. ACL.

Vempala, A., Blanco, E., & Palmer, A. (2018). Determining event durations: Models and error analysis. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 164–168. ACL.

Verhagen, M. (2004). Times between the lines. *Brandeis University, Massachusetts*.

Verhagen, M. (2005). Temporal closure in an annotation environment. In *Language Resources and Evaluation*, Vol. 39, pp. 211–241. Springer.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., & Pustejovsky, J. (2007). SemEval-2007 Task 15: TempEval temporal relation identification. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 75–80. ACL.

Verhagen, M., Knippen, R., Mani, I., & Pustejovsky, J. (2006). Annotation of temporal relations with Tango. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 2249–2252. ELRA.

Verhagen, M., & Pustejovsky, J. (2008). Temporal processing with the TARSQI toolkit. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pp. 189–192. ACL.

Verhagen, M., Sauri, R., Caselli, T., & Pustejovsky, J. (2010). SemEval-2010 Task 13: TempEval-2. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pp. 57–62. ACL.

Vilain, M., Kautz, H., & Van Beek, P. (1990). Constraint propagation algorithms for temporal reasoning: A revised report. In *Readings in Qualitative Reasoning about Physical Systems*, pp. 373–381. Elsevier.

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space and quantity. In *Trends in Cognitive Sciences*, Vol. 7, pp. 483–488. Elsevier.

Wang, Z., Shou, L., Chen, K., Chen, G., & Mehrotra, S. (2015). On summarization and timeline generation for evolutionary tweet streams. In *IEEE Transactions on Knowledge and Data Engineering*, Vol. 27, pp. 1301–1315. IEEE.

Williams, J., & Katz, G. (2012). Extracting and modeling durations for habits and events from Twitter. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 223–227. ACL.

Yan, R., Kong, L., Huang, C., Wan, X., Li, X., & Zhang, Y. (2011). Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 433–443. ACL.

Yoshikawa, K., Riedel, S., Asahara, M., & Matsumoto, Y. (2009). Jointly identifying temporal relations with Markov logic. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pp. 405–413. ACL.

Zhao, X. W., Guo, Y., Yan, R., He, Y., & Li, X. (2013). Timeline generation with social attention. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1061–1064. ACM.

Zhou, L., & Hripcsak, G. (2007). Temporal reasoning with medical dataa review with emphasis on medical natural language processing. In *Journal of Biomedical Informatics*, Vol. 40, pp. 183–202. Elsevier.