

A Survey on Temporal Reasoning for Temporal Information Extraction from Text (Extended Abstract)*

Artuur Leeuwenberg¹ and Marie-Francine Moens²

¹Julius Center, University Medical Center Utrecht, The Netherlands

²Department of Computer Science, KU Leuven, Belgium
aleeuw15@umcutrecht.nl, sien.moens@cs.kuleuven.be

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 the integration of symbolic reasoning with machine learning-based information extraction systems.

1 Introduction

Human language is filled with cues about the timing of the events that we communicate about. **Temporal Information Extraction (TIE)** is the process of automatically extracting temporal cues from text, with the goal to construct a (possibly underspecified) timeline of events from them.

TIE not only plays a major role in the general problem of natural language understanding (NLU), but is also used in many applications: information retrieval [Campos *et al.*, 2015], question answering [Meng *et al.*, 2017], and multi-document summarization [Ng *et al.*, 2014], and has great potential in the clinical domain, for applications like patient timeline visualization [Jung *et al.*, 2011], forecasting treatment effects [Zhou and Hripcsak, 2007], and patient selection for clinical trials [Raghavan *et al.*, 2014]. Because of the strong structure of time, and the great variation in the types of temporal cues we can express in language, a central challenge in TIE is how to combine all these separate cues into a single coherent temporal ordering of the described events. To achieve this, temporal reasoning is of vital importance. We define **Temporal Reasoning (TR)** in the context of TIE as the

*This paper is an extended abstract of an article in the Journal of Artificial Intelligence Research [Leeuwenberg and Moens, 2019].

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 all stages of TIE model construction: annotation, pre-processing, model training, inference, and evaluation.

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 this is discussed in depth by [Derczynski, 2017], and can be considered complementary to the focus of the survey. This extended abstract aims to capture the most important elements and conclusions from the original article [Leeuwenberg and Moens, 2019].

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

- An introduction to the various types of temporal information in language, and a brief explanation introducing the background on TR for TIE required for the latest state-of-the-art TIE models.
- An overview of the various ways in which TR has been exploited in TIE models over the past decades.
- A distillation of the most important conclusions to successfully incorporate TR in a TIE system and some promising directions for future work.

2 Temporal Cues in Language

It is important to study the types of temporal information that can be expressed in language because the different cues need to be combined by a TR system in order to build a complete temporal overview of the text. The most common interpretation of events on the timeline is as an interval (some expressions may refer to sets of events/intervals, or very short almost instantly occurring events). The span of the interval corresponds to the time that the event takes place. In this section, we give a short exemplified overview of different temporal cues in language to show what types of temporal information the cues can provide, i.e., in what way the cues may constrain the position of event intervals on the timeline.

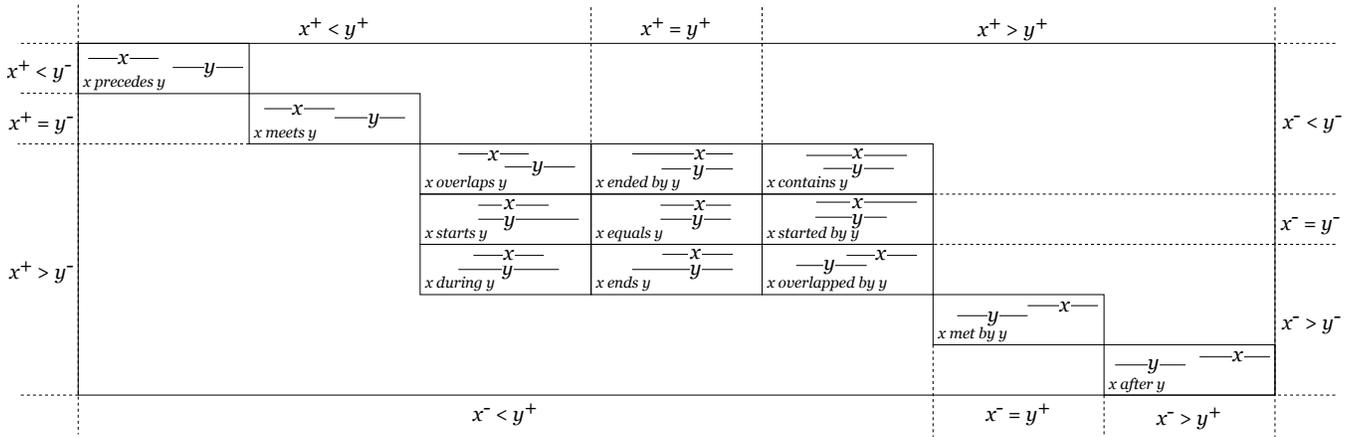


Figure 1: A lattice [Freksa, 1992] showing Allen’s 13 basic *interval* relations (small boxes), and their mapping to *point*-algebraic constraints (outside the large rectangle). Start and end points of intervals x and y are noted by x^- , x^+ , and y^- , y^+ respectively. Conceptual neighborhood between interval relations is proportional to Manhattan distance [Leeuwenberg and Moens, 2019].

Temporal cues can refer to various components of the event intervals: the entire *interval*, but also just its *start*, *end*, or *duration*. For instance, in Ex. 1, the *duration* of the antibiotics administration is given (*10 days*), and so is its *start time* (sometime on the 2nd of June). However, for the improvement in respiratory status only the *end time* is mentioned explicitly (last 2 days of the antibiotics administration).

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

The type of information given for each component can also differ: it can be *absolute* (e.g., in Ex. 1, the duration of the antibiotics of *10 days*), or *relative* to other intervals (e.g., in Ex. 2 the start time of the third set is *after* the first two). Additionally, the information can be *definite*, with clear interpretable timings (e.g., in Ex. 2, the duration of *28 minutes*), or indefinite, using *vague* quantification (in Ex. 2, the duration of the first two sets, each *nearly an hour*).

Ex. 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.*

Even with all these explicit cues, *world knowledge* about the typical order of events, their durations, and even typical time between events may be needed to properly interpret the temporal meaning that is to be conveyed (e.g., although not explicitly mentioned, one would assume that the third set took place on the same day as the first two). Finally, temporal information provided through language is generally *underspecified* (e.g., based on the example texts, we don’t know in what year any of the events took place). A good TR system should be able to flexibly incorporate all these different types of information, and possibly their uncertainty, to come to the timeline that is to be conveyed by the writer.

3 Reasoning with Intervals and Points

The most popular TR framework used in TIE is Allen’s interval algebra (AIA) [Allen, 1983]. AIA is based on a set of thirteen mutually exclusive basic interval relations (shown in Fig. 1) that can be assigned to any pair of *definite* intervals

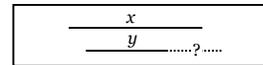


Figure 2: General Allen relation: x {contains, overlaps, ended by} y , described by the text “ y started during x ”.

(where relative positions of start and end-points are known). If for two intervals x and y information about the relative positions of their start and end-points is incomplete (i.e., their relation is indefinite, as in Fig. 2) a *general* Allen relation can be assigned, expressed as the disjunction of all possible basic relations, given the information provided. The full AIA covers the powerset of all thirteen basic relations, comprising $2^{13} = 8192$ relations.

To infer new relations from a graph of general Allen relations between different events, a composition table is used, containing 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. For the full AIA, computing all inferrable relations (taking the transitive closure) is, like checking consistency, NP-complete [Vilain *et al.*, 1990], making full AIA impractical for TIE systems. Many TIE systems that use TR are limited to using basic relations (i.e., definite relations). Although hardly used in TIE systems, many tractable sub-fragments of AIA exist [Krokhin *et al.*, 2003]. Some of which focus on integrating quantitative reasoning [Meiri, 1996] or uncertainty [Schockaert and De Cock, 2008]. A major insight on which many efficient TR algorithms are built is the fact that the basic Allen relations can be expressed as sets of point-algebraic relations ($<$, $=$, $>$), as shown in Fig. 1. Point algebra (PA) can express each basic Allen relation as a *conjunctive* set of point-wise relations between the starts and endings of the intervals. For example, x {equals} y can be expressed by the conjunctive set $\{x^- = y^-, x^+ = y^+\}$, i.e., if intervals x and y are equal, then their start and endings are also equal (and vice-versa). They also allow expression of many indefinite relations (e.g., the one from Fig. 2, as $\{x^- < y^-, x^+ > y^-\}$). Reasoning in PA requires a much smaller transitivity table, which contributes to its efficiency, important for practical TR systems.

4 Annotation Schemes

The scheme for temporal annotation in text steers the types of temporal cues that can be extracted by TIE systems, and consequently how TR can be exploited.

The most popular scheme is TimeML [Pustejovsky *et al.*, 2003], used as the basis for many corpora, and for TempEval, a series of shared tasks on TIE evaluation that resulted in many of the existing TIE systems [Verhagen *et al.*, 2007]. The core TimeML concepts are event expressions and time expressions (timex). Event expressions refer to events in the real world, and timex refer to calendar dates, times, and definite durations, among others. In TimeML, events and timex expressions are temporally connected through interval relations, called temporal links (TLinks), approximately following Allen’s basic interval relations, consequently focusing on definite information. To go beyond the focus on explicit and definite information in TimeML, many interesting proposals have been made: to include a *vague* relation to annotate indefinite relations [Cassidy *et al.*, 2014], to annotate point-wise relations between start and end points of events instead of basic interval relations [Ning *et al.*, 2018], annotation of durations for all events (not only for events with an explicit definite duration timex as in TimeML) [Pan *et al.*, 2006], and to annotate temporal position of events on the absolute calendar instead of by interval relations [Reimers *et al.*, 2016], as pair-wise temporal relation annotation is a difficult task for annotators [Derczynski, 2016]. These proposals provide new opportunities for further exploitation of TR.

5 Survey of Current Approaches

This section describes how TR has been incorporated in existing TIE systems throughout the steps of model development: annotation, data pre-processing, training and prediction, and evaluation. This abstract focuses purely on the task of temporal relation extraction among events and timex, as TR plays a large role in this task, and performance metrics are generally lower compared to event and timex extraction.

5.1 Annotation & Data Preprocessing

A widely recognized problem in temporal annotation, is that annotators regularly miss temporal cues in the text, especially when annotating TLinks between pairs of events, as there are so many possible relations to annotate. [Verhagen, 2005] argue that not all event-pairs are equally useful to annotate, as some can be inferred with TR from earlier annotations. Another motivation to exploit TR during annotation is that annotators sometimes annotate temporally inconsistent graphs, which impedes training of TIE systems.

A frequently made TR-based preprocessing step is to calculate the transitive closure over the annotated relations, to obtain a more dense graph. The effect of temporal closure on performance for models that use no TR during prediction is ambivalent. It may improve performance [Mani *et al.*, 2006], or decrease performance [Tatu and Srikanth, 2008]. However, when using a model that exploits TR also during prediction, data expansion through a temporal closure can be beneficial [Chambers and Jurafsky, 2008].

5.2 Training and Prediction

To enhance temporal relation extraction models, TR can be integrated during training, and prediction. There are various ways how integration of TR in training and prediction can improve TIE models: (1) ensure temporal consistency among predicted relations, and (2) to constrain the output space and improve prediction accuracy.

Greedy Inference. Two greedy approaches have been proposed to construct a graph of temporal relations: relations are added one-by-one, in a certain order, checking temporal consistency after each addition. The order can be natural reading order, or follow the score from a local classifier (best first). Both approaches improved results over absence of TR, the last surpassing the first [Bramsen *et al.*, 2006].

Post-Hoc Conflict Resolution. Another approach is to first predict the full graph with a local model, and detect and resolve inconsistencies afterwards, by removing conflicting edges based on confidence [Verhagen and Pustejovsky, 2008; Tatu and Srikanth, 2008] among others. Most works report a positive impact of this method, but differ in how they remove edges. It could be interesting to compare these to each other in more detail.

Sieve-Level and Stacked Inference. The sieve-based method [Chambers *et al.*, 2014] is flexible in incorporation of both rule-based and machine learning components. This approach extract TLinks in consecutive phases by separate model components (or sieves). Each sieve extracts TLinks, using the original input text, and the outputs from earlier sieves. TR is incorporated by taking a transitive closure after applying each sieve. This prevents later sieves from assigning TLinks that are inconsistent with those extracted by earlier sieves. Similar approaches have been adopted by [McDowell *et al.*, 2017], and [Mirza and Tonelli, 2016], who report an increase in recall (and F_1) due to sieve-level TR.

Much alike are methods that stack learners on top of each other to correct mistakes from earlier stages [Laokulrat *et al.*, 2015], which can even be trained jointly across stacks [Meng and Rumshisky, 2018].

Markov Logic Networks. Another important method for TR-based prediction are Markov Logic Networks (MLN), first explored for TIE by [Yoshikawa *et al.*, 2009]. A major difference between MLN and the methods mentioned earlier, is that instead of combining locally trained models in a global inference setting, MLN also incorporate the temporal constraints in training. The weights for TR constraints allow the model to learn soft correlations between TLinks. Model different transitivity rules with MLN, [Yoshikawa *et al.*, 2009] show that incorporating TR during training and prediction outperforms local models.

Integer Linear Programming. Where MLN use soft constraints, integer linear programming (ILP), a method for constrained optimization, uses hard constraints to cut off large areas from the search space, which can result in faster inference [Mojica and Ng, 2016]. ILP for TR-based prediction was first proposed by [Bramsen *et al.*, 2006], to maximize the sum of pair-wise classifier scores under TR-based constraints. ILP

has been adopted widely and successfully for TIE. In contrast to many, [Denis and Muller, 2011] formulated their ILP objective in PA for more efficient TR. Recent approaches not only use ILP during prediction, but also during training, combining ILP with a structured perceptron [Leeuwenberg and Moens, 2017; Ning *et al.*, 2017].

Direct Timeline Models. Recently, a new approach to temporal event ordering was proposed [Leeuwenberg and Moens, 2018]. Instead of predicting TLinks among events and temporal expressions, their model predicts start and end points of events such that their order satisfies the annotated TLinks. Prediction time of this approach is linear in the number of events, in contrast to predicting a graph of TLinks, which is quadratic or requires pruning. To train their model they exploit TR to convert TLinks to sets of PA constraints.

5.3 Evaluation

A problem in evaluation of TIE, is that the same timeline can be represented by different sets of relations: if a system predicts BEFORE(x, y) it should not be penalized if the ground truth is AFTER(y, x). For this, taking a temporal closure over the predicted and ground truth graphs has been proposed, in combination with different ways to weigh original vs. inferred relations [Setzer *et al.*, 2003; Tannier and Muller, 2011; UzZaman and Allen, 2011], of which the *temporal awareness* metric [UzZaman and Allen, 2011] is the most widely used evaluation metric.

6 Conclusions and Future Work

In this abstract, we provide a short demonstration of various types of temporal information in language, briefly introduced the most frequently used framework in TR for TIE, and reviewed different methods to exploit TR for constructing TIE models, distilling the most widely confirmed conclusions.

In closing, it is clear that TR is crucial for TIE, and widely used in all aspects of model construction: annotation, data preprocessing, training, prediction, and evaluation. 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) machine learning methods, to facilitate flexibility in model construction, multi-task learning, and sharing of representations.

Acknowledgements

This work was supported by the ERC Advanced Grant CALCULUS (788506), and by the IWT-SBO project ACCUMULATE (150056).

References

- [Allen, 1983] James F Allen. Maintaining knowledge about temporal intervals. In *Communications of the ACM*, pp. 832–843. ACM, 1983.
- [Bramsen *et al.*, 2006] Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. Inducing temporal graphs. In *Proc. EMNLP*, pp. 189–198. ACL, 2006.
- [Campos *et al.*, 2015] Ricardo Campos, Gaël Dias, Alípio M Jorge, and Adam Jatowt. Survey of temporal information retrieval and related applications. *ACM Computing Surveys*, 47(2):15, 2015.
- [Cassidy *et al.*, 2014] Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. An annotation framework for dense event ordering. In *Proc. ACL*, vol. 2, pp. 501–506. ACL, 2014.
- [Chambers and Jurafsky, 2008] Nathanael Chambers and Dan Jurafsky. Jointly combining implicit constraints improves temporal ordering. In *Proc. EMNLP*, pp. 698–706. ACL, 2008.
- [Chambers *et al.*, 2014] Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. Dense event ordering with a multi-pass architecture. In *Transactions of the ACL*, vol. 2, pp. 273–284, 2014.
- [Denis and Muller, 2011] Pascal Denis and Philippe Muller. Predicting globally-coherent temporal structures from texts via endpoint inference and graph decomposition. In *Proc. IJCAI*, pp. 1788–1793, 2011.
- [Derczynski, 2016] Leon Derczynski. Representation and learning of temporal relations. In *Proc. COLING*, pp. 1937–1948. ACL, 2016.
- [Derczynski, 2017] Leon RA Derczynski. *Automatically Ordering Events and Times in Text*. Springer, 2017.
- [Freksa, 1992] Christian Freksa. Temporal reasoning based on semi-intervals. In *Artificial Intelligence*, vol. 54, pp. 199–227. Elsevier, 1992.
- [Jung *et al.*, 2011] Hyuckchul Jung, James Allen, Nate Blaylock, Will De Beaumont, Lucian Galescu, and Mary Swift. Building timelines from narrative clinical records: initial results based-on deep natural language understanding. In *Proc. BioNLP*, pp. 146–154, 2011.
- [Krokhin *et al.*, 2003] Andrei Krokhin, Peter Jeavons, and Peter Jonsson. Reasoning about temporal relations: The tractable subalgebras of Allen’s interval algebra. In *Journal of the ACM*, vol. 50, pp. 591–640. ACM, 2003.
- [Laokulrat *et al.*, 2015] Natsuda Laokulrat, Makoto Miwa, and Yoshimasa Tsuruoka. Stacking approach to temporal relation classification with temporal inference. In *Journal of NLP*, vol. 22, pp. 171–196, 2015.
- [Leeuwenberg and Moens, 2017] Artuur Leeuwenberg and Marie-Francine Moens. Structured learning for temporal relation extraction from clinical records. In *Proc. EACL*, pp. 1150–1158. ACL, 2017.
- [Leeuwenberg and Moens, 2018] Artuur Leeuwenberg and Marie-Francine Moens. Temporal information extraction

- by predicting relative time-lines. In *Proc. EMNLP*, pp. 1237–1246, 2018.
- [Leeuwenberg and Moens, 2019] Artuur Leeuwenberg and Marie-Francine Moens. A survey on temporal reasoning for temporal information extraction from text. *Journal of Artificial Intelligence Research*, 66:341–380, 2019.
- [Mani *et al.*, 2006] Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. Machine learning of temporal relations. In *Proc. COLING-ACL*, pp. 753–760. ACL, 2006.
- [McDowell *et al.*, 2017] Bill McDowell, Nathanael Chambers, Alexander Ororbia II, and David Reitter. Event ordering with a generalized model for sieve prediction ranking. In *Proc. IJCNLP*, vol. 1, pp. 843–853. ACL, 2017.
- [Meiri, 1996] Itay Meiri. Combining qualitative and quantitative constraints in temporal reasoning. In *Artificial Intelligence*, vol. 87, pp. 343–385, 1996.
- [Meng and Rumshisky, 2018] Yuanliang Meng and Anna Rumshisky. Context-aware neural model for temporal information extraction. In *Proc. ACL*, vol. 1, pp. 527–536. ACL, 2018.
- [Meng *et al.*, 2017] Yuanliang Meng, Anna Rumshisky, and Alexey Romanov. Temporal information extraction for question answering using syntactic dependencies in an LSTM-based architecture. In *Proc. EMNLP*, pp. 887–896. ACL, 2017.
- [Mirza and Tonelli, 2016] Paramita Mirza and Sara Tonelli. Catena: Causal and temporal relation extraction from natural language texts. In *Proc. COLING*, pp. 64–75. ACL, 2016.
- [Mojica and Ng, 2016] Luis Gerardo Mojica and Vincent Ng. Markov Logic Networks for text mining: A qualitative and empirical comparison with ILP. In *Proc. LREC*, pp. 4388–4395, 2016.
- [Ng *et al.*, 2014] Jun-Ping Ng, Yan Chen, Min-Yen Kan, and Zhoujun Li. Exploiting timelines to enhance multi-document summarization. In *Proc. ACL*, vol. 1, pp. 923–933. ACL, 2014.
- [Ning *et al.*, 2017] Qiang Ning, Zhili Feng, and Dan Roth. A structured learning approach to temporal relation extraction. In *Proc. EMNLP*, pp. 1027–1037. ACL, 2017.
- [Ning *et al.*, 2018] Qiang Ning, Hao Wu, and Dan Roth. A multi-axis annotation scheme for event temporal relations. In *Proc. ACL*, pp. 1318–1328, 2018.
- [Pan *et al.*, 2006] Feng Pan, Rutu Mulkar-Mehta, and Jerry R Hobbs. An annotated corpus of typical durations of events. In *Proc. LREC*, pp. 77–82, 2006.
- [Pustejovsky *et al.*, 2003] James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. TimeML: Robust specification of event and temporal expressions in text. In *New Directions in QA*, vol. 3, pp. 28–34, 2003.
- [Raghavan *et al.*, 2014] Preethi Raghavan, James L Chen, Eric Fosler-Lussier, and Albert M Lai. How essential are unstructured clinical narratives and information fusion to clinical trial recruitment? In *AMIA Jt. Summits Transl. Sci. Proc.*, pp. 218–223, 2014.
- [Reimers *et al.*, 2016] Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. Temporal anchoring of events for the TimeBank corpus. In *Proc. ACL*, vol. 1, pp. 2195–2204, 2016.
- [Schockaert and De Cock, 2008] Steven Schockaert and Martine De Cock. Temporal reasoning about fuzzy intervals. In *Artificial Intelligence*, vol. 172, pp. 1158–1193. Elsevier, 2008.
- [Setzer *et al.*, 2003] Andrea Setzer, Robert Gaizauskas, and Mark Hepple. Using semantic inferences for temporal annotation comparison. In *Proc. IWCS*, 2003.
- [Tannier and Muller, 2011] Xavier Tannier and Philippe Muller. Evaluating temporal graphs built from texts via transitive reduction. In *Journal of Artificial Intelligence Research*, vol. 40, pp. 375–413, 2011.
- [Tatu and Srikanth, 2008] Marta Tatu and Munirathnam Srikanth. Experiments with reasoning for temporal relations between events. In *Proc. COLING*, pp. 857–864. ACL, 2008.
- [UzZaman and Allen, 2011] Naushad UzZaman and James F. Allen. Temporal evaluation. In *Proc. ACL*, pp. 351–356, 2011.
- [Verhagen and Pustejovsky, 2008] Marc Verhagen and James Pustejovsky. Temporal processing with the TARSQI toolkit. In *Proc. COLING*, pp. 189–192. ACL, 2008.
- [Verhagen *et al.*, 2007] Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. SemEval-2007 Task 15: TempEval temporal relation identification. In *Proc. SemEval*, pp. 75–80. ACL, 2007.
- [Verhagen, 2005] Marc Verhagen. Temporal closure in an annotation environment. In *Language Resources and Evaluation*, vol. 39, pp. 211–241, 2005.
- [Vilain *et al.*, 1990] Marc Vilain, Henry Kautz, and Peter Van Beek. Constraint propagation algorithms for temporal reasoning: A revised report. In *Readings in Qualitative Reasoning about Physical Systems*, pp. 373–381. Elsevier, 1990.
- [Yoshikawa *et al.*, 2009] Katsumasa Yoshikawa, Sebastian Riedel, Masayuki Asahara, and Yuji Matsumoto. Jointly identifying temporal relations with Markov logic. In *Proc. ACL-IJCNLP*, pp. 405–413, 2009.
- [Zhou and Hripcsak, 2007] Li Zhou and George Hripcsak. Temporal reasoning with medical data - a review with emphasis on medical natural language processing. In *J. Biomed. Inform.*, vol. 40, pp. 183–202, 2007.