Transformer-Based Temporal Information Extraction and Application: A Review

Xin Su
Intel
xin.su@intel.com

Phillip Howard
Thoughtworks
phillip.howard@thoughtworks.com

Steven Bethard
University of Arizona
bethard@arizona.edu

Abstract

Temporal information extraction (IE) aims to extract structured temporal information from unstructured text, thereby uncovering the implicit timelines within. This technique is applied across domains such as healthcare, newswire, and intelligence analysis, aiding models in these areas to perform temporal reasoning and enabling human users to grasp the temporal structure of text. Transformer-based pre-trained language models have produced revolutionary advancements in natural language processing, demonstrating exceptional performance across a multitude of tasks. Despite the achievements garnered by Transformer-based approaches in temporal IE, there is a lack of comprehensive reviews on these endeavors. In this paper, we aim to bridge this gap by systematically summarizing and analyzing the body of work on temporal IE using Transformers while highlighting potential future research directions.

1 Introduction

Temporal information extraction (IE) is a critical task in natural language processing (NLP). Its objective is to extract structured temporal information from unstructured text, thereby revealing the implicit timelines within the text. This not only helps improve temporal reasoning in other NLP tasks, such as timeline summarization and temporal question answering, but also helps human users in gaining a deeper understanding of the evolution of text content over time. For example, Figure 2 displays a snippet of George Washington's Wikipedia page and the timeline of his position changes; relying solely on text-heavy documents to trace his position changes over different years is time-consuming and may lack accuracy as facts and temporal expressions are scattered throughout the text. In contrast, a timeline enables both NLP models and humans to understand the changes in these positions over time more succinctly and clearly. The application of this structured temporal information is not limited to Wikipedia but is also widely used in other domains such as healthcare (Styler IV et al., 2014).

The advent of the Transformer architecture (Vaswani et al., 2017) has sparked a revolutionary change in the field of NLP, particularly with the recent Transformer-based generative large language models (LLM), such as LLAMA3 (Dubey et al., 2024) and GPT-4 (Achiam et al., 2023), demonstrating exceptional performance across many tasks. Nevertheless, there has yet to be an in-depth study that provides a comprehensive review or analysis of the Transformer architecture's application in the field of temporal IE. Existing surveys (Lim et al., 2019; Leeuwenberg and Moens, 2019; Alfattni et al., 2020; Olex and McInnes, 2021) focus on rule-based systems or traditional machine learning models (e.g., support vector machines) which are reliant on hand-crafted features. Only Olex and McInnes (2021) touches on the application of Transformer models, but they offer only a brief description of BERT-style models and focus largely on the clinical domain.

To address this gap, we systematically review the applications of Transformer-based models in the field of temporal IE. Broadly, temporal IE refers to any tasks involving the extraction of temporal information from text. We focus on three important tasks which are defined in the most widely adopted temporal IE annotation framework, TimeML (Pustejovsky, 2003): time expression identification, time expression normalization, and temporal relation extraction. Our contributions are summarized as follows: (1) We systematically review, summarize, and categorize the existing temporal IE datasets, Transformer-based methods, and applications. (2) We identify and highlight the research gaps in the field of temporal IE and suggest potential directions for future research.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28822–28841
November 4-9, 2025

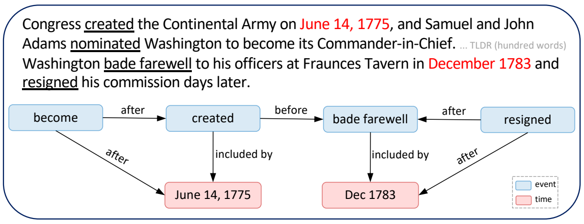

Figure 1: A snippet from George Washington's Wikipedia page and the corresponding temporal graph.

2 Overview

The goal of temporal IE is to extract structured temporal information from unstructured text, facilitating its interpretation and processing by computers, thereby achieving a transformation from text to structure. The final result of a temporal IE system is the construction of a directed acyclic graph, or a temporal graph, which represents the structured temporal information in the text. In the temporal graph, nodes represent time expressions and events (temporal entities), while edges depict the temporal relations between these nodes, such as "before," "after," etc. For instance, Figure 1 illustrates a text snippet from George Washington's Wikipedia page and its corresponding temporal graph.

Constructing a temporal graph involves several sub-tasks: time expression identification, time expression normalization, event extraction, and temporal relation extraction. The following is a brief introduction to these sub-tasks; see Appendix B for a discussion of common evaluation methods.

Time Expression Identification and Normalization Time expression identification refers to identifying specific time points, durations, or periods within the text, such as the explicitly dateable expression "February 25, 2024," or more ambiguous expressions like "three days ago" (Pustejovsky, 2003). Time normalization involves converting identified expressions into a standardized format to improve their interpretability. For example, under the ISO-TimeML framework (Pustejovsky et al., 2010), "February 25, 2024" might be converted into the TIMEX3 format as "2024-02-25".

Event Trigger Extraction In temporal IE, event extraction differs from other NLP event extraction tasks; it simply marks the event trigger words that represent actions, such as "accident" in "about two weeks after the accident occurred". We will not review event extraction works because, to our knowledge, there is currently no temporal IE research focused solely on event extraction. Furthermore, most existing work on temporal IE assumes that event triggers have already been identified. For a comprehensive survey of event extraction, we refer readers to (Li et al., 2022).

Temporal Relation Extraction The task of temporal relation extraction aims to identify the temporal relations among given events and time expressions. Common temporal relations include before, after, and simultaneous. For example, in Figure 1, the temporal relation between "June 14, 1775" and the event "become" is marked as "after".

3 Datasets

A clearly defined annotation framework is essential when constructing a dataset for temporal IE. It needs to precisely define time expressions, events, and their relations. We summarize all the datasets in Table 1 of Appendix C.

3.1 TimeML Annotation Framework Datasets

An end-to-end temporal IE dataset encompasses various tasks, including the identification and normalization of time expressions and the extraction of temporal relations. Most end-to-end temporal information datasets have been based on the TimeML framework (Pustejovsky, 2003) or its derivatives, such as ISO-TimeML (Pustejovsky et al., 2010). We present datasets based on the TimeML framework in the first section of Table 1.

TimeBank (Pustejovsky, 2003) was the first dataset to adopt the TimeML framework, focusing on the English news domain. Follow-up works included the TempEval shared task series (Verhagen et al., 2007, 2010; UzZaman et al., 2013), covering multiple languages, including Chinese, English, Italian, French, Korean, and Spanish. There are also language-specific datasets like French TimeBank (Bittar et al., 2011), Spanish TimeBank (Nieto et al., 2011), Portuguese TimeBank (Costa and Branco, 2012), Japanese TimeBank (Asahara et al., 2013), Italian TimeBank (Bracchi et al., 2016), and Korean TimeBank (Lim et al., 2018). Similarly, the MeanTime dataset (Minard et al., 2016) offers data in English, Italian, Spanish, and Dutch. Datasets based on TimeML and its variants showcase language diversity and also cover several different domains: the Spanish TimeBank focuses on history text, the Korean TimeBank is based on Wikipedia content, and the Richer Event Description dataset

(O'Gorman et al., 2016) provides data from both news and forum discussion domains.

Additionally, efforts have been made to improve the temporal relation annotations in the original TimeBank. TimeBank-Dense (Chambers et al., 2014) addresses the sparsity of temporal relation annotations in TimeBank by requiring annotators to label all temporal relations within a given scope, thus increasing the number of temporal relations in the dataset. The TORDER dataset (Cheng and Miyao, 2018) annotates the same documents as TimeBank-Dense, introducing temporal relations automatically by anchoring times and events to absolute points, reducing the annotation burden. The MATRES dataset (Ning et al., 2018) focuses on events from TimeBank-Dense, anchoring events to different timelines and comparing their start times to enhance inter-annotator consistency.

Several datasets have been developed specific to the clinical domain, of which the Thyme datasets (Bethard et al., 2015, 2016, 2017) are most notable. They are based on the Thyme-TimeML (Styler IV et al., 2014) annotation framework, which adjusts and adds new temporal attributes from ISO-TimeML to suit medical texts. Like the TimeBank series, the Thyme dataset involves identifying and normalizing time expressions and extracting temporal relations, focusing on English. Another similar dataset is i2b2-2012 (Sun et al., 2013), which adapts the TimeML framework for clinical texts.

Besides end-to-end datasets, several others based on TimeML or its variants focus on specific temporal IE tasks. For instance, the AncientTimes dataset (Strötgen et al., 2014) covers a broad range of languages, concentrating on the identification and normalization of time expressions. The TD-Discourse dataset (Naik et al., 2019), based on TimeBank-Dense, expands the annotation window for temporal relations, focusing on their extraction. The German time expression (Strötgen et al., 2018) and German VTEs (May et al., 2021) datasets are dedicated to identifying and normalizing time expressions in German. The PATE dataset (Zarcone et al., 2020) provides data aimed at time expression identification and normalization for the virtual assistant domain.

3.2 Other Annotation Framework Datasets

Unlike datasets for temporal IE based on TimeML, other annotation frameworks typically focus on specific sub-tasks of temporal IE, such as time expression identification and normalization or the extraction of temporal relations. We present these datasets in the second section of Table 1.

For time expression identification and normalization, WikiWars (Mazur and Dale, 2010) and SCATE (Laparra et al., 2018) are two major datasets. WikiWars contains data from English and German Wikipedia, annotated based on TIMEX2 (a precursor to TimeML's TIMEX3) to mark explicit time expressions. The SCATE dataset, based on English news and clinical documents, aims to address limitations in TimeML that prevent expressing multiple calendar units, times relative to events, and compositional time expressions. To achieve this, SCATE represents time expressions as compositions of temporal operators.

For temporal relations, there are datasets based on the temporal dependency tree/graph (Zhang and Xue, 2018, 2019; Yao et al., 2020) and CaTeRS (Mostafazadeh et al., 2016) frameworks. Unlike the pairwise temporal relations considered in the TimeML framework, temporal dependency tree assumes that all time expressions and events in a document have a reference time, allowing for the representation of overall temporal relations through a dependency tree. The subsequent temporal dependency graph dataset (Yao et al., 2020) relaxed this assumption by enabling each event in a document to have a reference event, a reference time, or both, thus forming a temporal graph structure. The temporal dependency tree dataset covers news and narrative domains in English and Chinese, while the temporal dependency graph dataset focuses on English news. Meanwhile, CaTeRS concentrates on analyzing temporal relations between events in English commonsense stories, with event definitions based on ontologies, different from the verb-, adjective-, or noun-based definitions in TimeML. CaTeRS' annotation of temporal relations is story-wide, with a simplified set of relations. We present additional timeline focused datasets at Appendix D.

3.3 Discussion and Research Gaps

Domain Bias Existing annotated datasets exhibit significant domain biases. As demonstrated in Table 1, among the 32 datasets we reviewed, 20 (or 63%) are predominantly focused on the newswire domain. While temporal information is crucial for understanding news content, an excessive concentration in a single domain hampers the advancement and generalizability of systems trained on these datasets, since the challenges and difficulties encountered in temporal IE vary across different domains. Notably, the Clinical TempEval 2017 shared task (Bethard et al., 2017) reveals that most tasks suffer an approximately 20-point drop in performance in a cross-domain setting, underscoring how domain shifts can significantly degrade model accuracy. For example, temporal information, especially time expressions, in newswire texts tend to be explicitly stated, whereas in other domains, like historical Wikipedia entries, they might appear in subtler ways. Consider a statement from a page about George Washington that reads, "... 1798, one year after that, he stepped down from the presidency," which would demand a more nuanced interpretation for accurate time normalization. Cultivating datasets that represent a variety of domains is vital to driving innovation in temporal IE.

Language Diversity Unlike the domain homogeneity of the datasets, the existing datasets display rich linguistic diversity, covering 15 different languages. The representation of time varies across languages, and even when semantically similar, the specific time intervals on the timeline can differ. For example, analysis in Shwartz (2022) shows that different cultures/languages have significant variations in the understanding of "night" and "evening" during the day. One instance is that Brazilian Portuguese speakers often use "evening" and "night" interchangeably to denote the same time period, possibly because the tropical climate in Brazil causes evening to transition quickly into night. However, this might not be applicable to other cultures or languages. Therefore, the language diversity in datasets is crucial for developing models capable of effectively extracting temporal information across different languages.

Annotation and Dataset Framework Development Slows Down Aside from the original TimeML and some incremental modifications to it, no new end-to-end temporal IE annotation frameworks have been proposed. A significant issue with the existing TimeML-based annotation frameworks is the limited amount of information that the resultant temporal graphs can represent. For instance, in Figure 1, we only see trigger words for events, time expressions, and some temporal relations. When these temporal graphs are isolated from their original context and treated as stand-alone entities, they struggle to provide a comprehensive understanding of the textual information. This might explain why, in the upcoming Section 6, we see no work directly employing these extracted temporal graphs for reasoning to accomplish specific tasks, such as answering temporal questions. Instead, these temporal graphs are used as auxiliary tools or additional knowledge to assist task-specific models in temporal reasoning.

In addition to the stagnation in the innovation of end-to-end annotation frameworks, there has been a notable decline in dataset development efforts in the field of temporal IE in recent years. This trend may primarily stem from the intrinsic complexity of the annotation process for temporal IE datasets. Such complexity accounts for the low annotator agreement observed in many annotation tasks (Cassidy et al., 2014). Furthermore, as demonstrated by analysis in Su et al. (2021), even Ph.D. students in relevant fields find it challenging to comprehend annotation guidelines and annotate high-quality data within a short period. These issues highlight the difficulties in developing temporal IE datasets, suggesting that improvements in the annotation framework might be necessary to address these challenges.

4 Time Expression Methods

4.1 Methods Overview

In the realm of time expression identification, most prior work (Almasian et al., 2021; Chen et al., 2019; Mirzababaei et al., 2022; Olex and McInnes, 2022; Laparra et al., 2021; Almasian et al., 2022; Cao et al., 2022) leverages discriminative models built upon Transformer encoders like BERT (Devlin et al., 2019). These approaches typically frame time expression identification as a token classification task, wherein a sequence of tokens is input, processed through a base encoder model to obtain contextualized representations, and these representations are fed into a classifier (such as a simple linear classification layer or a Conditional Random Field layer) to identify time expressions and their specific types. Almasian et al. (2021) is the only work exploring a generative approach for time expression identification, framing the task as a sequence-to-sequence problem and employing a pair of Transformer encoders to formulate an encoder-decoder model—where one serves as the encoder and the other as the decoder—to generate additional TIMEX3 tags for the input, thereby recognizing time expressions and their types.

Shwartz (2022) and Kim et al. (2020) focus on the normalization of time expressions and use Transformer-based models. Shwartz (2022) aims to normalize time expressions from various cultural contexts (e.g., morning, noon, afternoon) into precise hourly representations within a day. They train a BERT model with a masked language modeling task to predict specific times of day that are masked, given the time expressions. Kim et al. (2020) seeks to normalize time expressions in novels into specific daily hours, fine-tuning the BERT model for a 24-class classification task to ascertain the corresponding times of day for given expressions.

Lange et al. (2023) addresses both extraction and normalization of time expressions, adopting a pipeline approach. Initially, they fine-tune the XLM-R model using the token classification method to extract time expressions, then denote identified expressions with TIMEX3 tags with masked time values, and finally fine-tune the XLM-R model with masked language modeling to predict the normalized masked time values.

Several of the aforementioned works also utilize data augmentation techniques to improve the model's multilingual performance (Lange et al., 2023; Mirzababaei et al., 2022; Almasian et al., 2022). For instance, Lange et al. (2023) employs the rule-based HeidelTime method (Strötgen and Gertz, 2010) to annotate time expressions and their normalizations across 87 languages, generating a semi-supervised dataset to facilitate model training.

4.2 Discussion and Research Gaps

Despite the significant achievements of Transformer models in various NLP tasks, research in the area of time expression identification and normalization has remained relatively limited over the past few years. This is particularly true of time normalization, where the volume and depth of research are low, especially when compared to similar tasks such as named entity recognition, entity normalization, and entity linking. Furthermore, the methodological diversity in existing works is notably constrained, with most research relying on pre-trained Transformer models for simple token classification. While generative LLMs like GPT-4 or LLAMA3 have demonstrated impressive performance in other NLP tasks, their potential in the identification and normalization of time expressions has barely been explored. This suggests a significant research gap exists; exploration of generative approaches may offer the potential for advancement in time expression identification and normalization.

5 Temporal Relation Methods

The task of temporal relation extraction typically assumes that events and time expressions in the text have already been identified, with the only objective being to extract the temporal relations between them. We summarize all the reviewed temporal relation extraction works in Appendix E Table 2. Discriminative methods typically employ a pretrained discriminative language model like BERT or RoBERTa (Liu et al., 2019) as the base encoder model to derive contextualized representations of events or time expressions. Subsequently, these representations are paired and input into a classification layer for a multi-class classification task, with each class representing a different temporal relation. Generative methods typically leverage encoder-decoder models such as T5 (Raffel et al., 2020) or decoder-only models like GPT (Radford et al., 2019) to generate a target sequence that encapsulates the temporal relation between the input events and times. These methods often rely on post-processing techniques to extract specific temporal relations from the predicted target sequences.

5.1 Discriminative Methods Overview

Works on discriminative temporal relation extraction have mainly focused on integrating external knowledge and improving model robustness.

5.1.1 Integrating External Knowledge

Commonsense Knowledge Commonsense knowledge for temporal relations usually involves typical sequences of events, such as eating typically occurring after cooking. Such commonsense knowledge might be fundamental for humans, but absent from the base encoder model. Ning et al. (2019), Wang et al. (2020) and Tan et al. (2023) integrated knowledge from external commonsense knowledge graphs. Tan et al. (2023) employs a complex Bayesian learning method to merge the knowledge with the contextualized representations from the base encoder, whereas Ning et al. (2019) and Wang et al. (2020) simply concatenate the vectorized representations of the commonsense knowledge with those from the base encoder.

Syntactic and Semantic Knowledge Syntactic and semantic knowledge, typically extracted using off-the-shelf external tools or straightforward rules, enrich the base encoder models' representations. For instance, Wang et al. (2022) utilizes SpaCy's dependency parser to parse the syntactic dependency trees from the input text and neuralcoref to identify coreferential relationships among entities. Mathur et al. (2021) employs the discoursegraphs library to parse rhetorical dependency graphs from the text. To integrate this structured knowledge into the contextualized event or time expression representations, graph neural networks are often employed over syntactic or semantic pairwise relations (Wang et al., 2022; Mathur et al., 2022; Zhou et al., 2022; Mathur et al., 2021). For example, Wang et al. (2022) first encodes an input sequence containing event pairs with the RoBERTa model to generate initial contextual representations, which are then enhanced with extracted syntactic and semantic knowledge using additional graph neural network layers. Another method is to prelearn or extract vectorized representations of the knowledge, which are later concatenated with the event or time expression representations (Ross et al., 2020; Wang et al., 2020; Han et al., 2019a; Ning et al., 2019; Han et al., 2019b; Yao et al., 2024a), as in Wang et al. (2020), where RoBERTa token embeddings and one-hot vectors of part-of-speech tags are combined.

Temporal-Specific Rules These rules are intrinsic to temporal relations themselves, with symmetry and transitivity being the most common. For instance, if event A happens before event B, then symmetry can be used to infer that B happens after A. And if A precedes B and B precedes C, transitivity can be used to infer that A precedes C. Detailed explanations of the symmetry and transitivity rules and a comprehensive transitivity table are provided in Ning et al. (2019). Recent works have incorporated these rules during both training and inference. During training, models employ various approaches including box embedding (Hwang et al., 2022), hyperbolic embedding (Tan et al., 2021), loss function regularization (Zhou et al., 2021; Wang et al., 2020), contrastive objectives (Niu et al., 2024), logical expressions over event time points (Huang et al., 2023), and hierarchical logical conditions (Ning et al., 2024). For inference, methods include custom heuristics (Wang et al., 2022; Zhou et al., 2022, 2021; Liu et al., 2021), linear programming formulation (Wang et al., 2020; Han et al., 2019c), and structured prediction with support vector machines (Han et al., 2019a).

Label Distribution Knowledge of label distribution pertains to the frequency distribution of specific temporal relations in the training set. Wang et al. (2023) and Han et al. (2020) integrate this distribution knowledge into their frameworks, using it as a regularization term in the loss function or for inference-time linear programming, aiming to mitigate potential biases in model predictions.

5.1.2 Improving Model Robustness

Multitask Learning Wang et al. (2022), Lin et al. (2020) and Cheng et al. (2020) categorize temporal relations and treat the extraction of different types of temporal relations as independent tasks, employing multitask learning to extract all types of relations simultaneously. For instance, Wang et al. (2022) delineates tasks into event-event, event-time, and event-document creation time, undergoing multitask training across these three tasks. Mathur et al. (2022) applies multitask learning in their model to concurrently predict temporal relations and dependency links between nodes in a temporal dependency tree. Similarly, Ballesteros et al. (2020) implements multitask learning by integrating the extraction of temporal relations with the extraction of entity relations in the general domain.

Data Augmentation Wang et al. (2023) generates counterfactual instances from the training set samples to mitigate model bias, while Tiesen and Lishuang (2022) employs predefined templates to create additional training examples.

Continued Pre-training of Base Encoder In Zhao et al. (2021) and Han et al. (2021), heuristic methods are used to identify temporal indicators in a corpus of unlabeled data, further training the base encoder using a masked language modeling (MLM) approach to recover masked indicators. Lin et al. (2019) focuses on the medical domain, using MLM on electronic health records from MIMIC-III to adapt the base encoder for domain-specific training prior to temporal relation extraction.

Adversarial Training Kanashiro Pereira (2022) and Pereira et al. (2021) introduce adversarial perturbations at different layers of the Transformer encoder during training to enhance model robustness.

Self-training Cao et al. (2021) and Ballesteros et al. (2020) initially train a temporal relation extraction model on annotated datasets and then apply the model to unlabeled data to obtain modelgenerated labels as pseudo labels. They subsequently select pseudo-labeled examples as sliver examples based on the model's uncertainty scores and confidence scores (probability scores for specific temporal relation predictions) to train the model.

5.2 Generative Methods Overview

Generative approaches in Temporal IE fall into two main categories: fine-tuned encoder-decoder models and large language model (LLM) prompting methods. For fine-tuned generative models, Dligach et al. (2022) investigate BART (Lewis et al., 2020) and T5 (Raffel et al., 2020) architectures, finding that producing outputs for each temporal entity pair separately outperforms triplet format (entity, relation, entity). Recent work has also explored LLM-based approaches. Yuan et al. (2023) and Huang et al. (2023) examine various prompting strategies, with Huang et al. (2023) demonstrating that structured, logic-informed prompts significantly improve performance over standard prompting. Hu et al. (2025) formulates temporal relation extraction as a question-answering task with rationale generation that includes coreference and transitive chains. Meanwhile, Niu et al. (2024) integrates LLMs specifically to enhance commonsense reasoning in their hybrid system. More recently, Eirew et al. (2025) address the computational inefficiency of pairwise classification by proposing a zero-shot method that generates a document's complete temporal graph in a single inference step. They employ temporal constraint optimization with Integer Linear Programming to ensure global consistency across relations, and introduce OmniTemp, a dataset with complete temporal relation annotations for all event pairs within documents. Despite these advances, current findings indicate that prompting-only approaches still underperform compared to fine-tuned discriminative models.

5.3 Discussion and Research Gaps

Homogenization of Methods and Evaluations While numerous Transformer-based methods for temporal relation extraction have emerged, they tend to be algorithmically similar, utilizing discriminative base models like BERT to represent temporal entities and incorporating additional knowledge into these representations. A common strategy involves using off-the-shelf IE tools to extract syntactic knowledge and enhance the base model's representations with graph neural networks. The small gains in state-of-the-art performance from one model to the next probably represent additional hyperparameter tuning more than substantial progress in understanding the relations between temporal entities in text.

Most works also focus on only three datasets – MATRES, TimeBank-Dense, and TDDiscourse – which are predominantly in the newswire domain with only 274, 36, and 34 documents, respectively, and exhibit significant overlap. This limitation in datasets might lead to an incomplete assessment of the models' generalization capabilities. Repeated testing and fine-tuning on these small, overlapping datasets could result in overfitting, failing to reflect the models' effectiveness on broader and more diverse datasets. Moreover, this singular domain-focused evaluation approach could cause severe domain bias, leaving the applicability of these methods outside the news domain uncertain. For a detailed comparative analysis of different methodological approaches and their trade-offs, see Appendix H.

Generative LLMs: Progress and Challenges Despite increasing interest in generative LLMs for temporal relation extraction, a significant research gap remains: current generative approaches consistently underperform compared to fine-tuned discriminative models (Yuan et al., 2023). Although recent works have explored structured prompts (Huang et al., 2023), question-answering frameworks (Hu et al., 2025), and hybrid systems (Niu et al., 2024), none have matched state-of-the-art discriminative methods. Promising directions for future research include: (1) specialized temporal fine-tuning techniques for LLMs; (2) more effective methods to encode temporal rules and constraints in LLM prompts; and (3) improved evaluation frameworks for generative outputs in temporal tasks.

Increased Demand for Model Openness As shown in the last column of Table 2, most temporal relation extraction models are not publicly available, possibly due to the absence of code releases or the need to re-train models on new datasets even when code is provided. Re-training a model involves significant replication work. This inaccessibility directly impacts the practical application and testing of these trained models in other temporal reasoning tasks, thereby affecting the development of the temporal relation extraction field. Given the application-oriented nature of temporal relation extraction tasks, only by understanding the specific issues encountered in actual applications can we propose strategies to address these real-world challenges.

6 Applications

6.1 Methods Overview

Temporal IE is often regarded as an "upstream" system, akin to other general IE systems. These systems aim to extract structured information to improve the reasoning of "downstream" tasks, such as temporal reasoning. A natural question is how the models from Sections 4 and 5 are used in downstream tasks to help temporal reasoning.

Despite a wealth of research on Transformer-based temporal IE systems in recent years, there has been scant application of these systems' outputs in temporal reasoning tasks. Only a few temporal reasoning tasks, such as timeline extraction, timeline summarization and temporal question answering, leverage the results of temporal IE. Timeline extraction is a direct product of temporal IE, where the extracted events and time expressions, along with their temporal relations, naturally form a chronologically ordered timeline following the traditional TimeML paradigm. For example, the recent Chemotherapy Timeline Extraction shared task (Yao et al., 2024b) focuses on constructing patient-level treatment timelines from electronic health records, with most participating systems using fine-tuned Transformer models for event and time expression extraction, followed by temporal relation classification. The timeline summarization task aims to chronologically order and label key dates of events within a collection of news documents, while temporal question answering relies on unstructured context documents to answer temporal-related questions. Both tasks require reasoning about time and events to generate outcomes.

One approach to utilizing temporal IE systems is to explicitly construct temporal graphs to assist with temporal reasoning. Some works use only simple temporal graphs containing only time expressions extracted by rules (Su et al., 2023) or Transformers (Yang et al., 2023; Xiong et al., 2024) and normalized by rules. Other works use complete temporal graphs constructed by a complete temporal IE pipeline, including time expression identification, normalization, and temporal relation extraction, with Mathur et al. (2022) using Transformer-based relation extraction, and Li et al. (2021) using LSTM-based relation extraction and rules for the other components. As for the usage of the constructed temporal graph, they can be input into models directly in text form (Su et al., 2023; Yang et al., 2023; Xiong et al., 2024) or encoded into the hidden states of a Transformer model through an attention fusion mechanism or graph neural networks (Li et al., 2021; Mathur et al., 2022; Su et al., 2023).

Some works only preprocess the input with a specific temporal IE component rather than building a temporal graph. For instance, Bedi et al. (2021) employs the rule-based HeidelTime (Strötgen and Gertz, 2010) for extracting and normalizing time expressions in texts for constructing the input of a temporal question generation model; while Cole et al. (2023) uses the rule-based SUTime (Chang and Manning, 2012) to process the entire Wikipedia, supporting the temporal pre-training of the Transformer model.

6.2 Discussion and Research Gaps

Although there is considerable work on Transformer-based temporal IE, especially in temporal relation extraction tasks, these methods have not been widely applied to downstream tasks. For example, there are many Transformer-based works that have been trained on the MATRES dataset, but none have been utilized in downstream tasks. This may be attributed to most temporal IE models not being publicly available, as shown in Table 2. Replicating these models can be both complex and time-consuming, requiring substantial effort. Furthermore, existing models exhibit domain bias. For example, in temporal relation extraction tasks, most research relies on the TimeBank-Dense and MATRES datasets, which primarily contain data from the newswire domain. Hence, the generalization capabilities of these models in other domains might be limited.

7 Conclusion

In this paper, we provide an overview of three classic tasks in the field of temporal IE: time expression identification, time expression normalization, and temporal relation extraction. We discuss datasets, Transformer-based methods, and their applications within these areas. We found that although Transformer models have demonstrated outstanding performance on many NLP tasks, there remain significant research gaps in the domain of temporal IE. We hope this survey will offer a comprehensive review and insights to researchers in the field, inspiring further research to address these existing gaps. We expand on the research opportunities arising from these gaps in Appendix F.

Limitations

In this review, we focus exclusively on Transformer-based temporal IE methods, without including rule-based approaches. We also center our discussion on the most common temporal IE tasks rather than addressing every possible subtask.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ghada Alfattni, Niels Peek, and Goran Nenadic. 2020. Extraction of temporal relations from clinical free text: A systematic review of current approaches. *Journal of biomedical informatics*, 108:103488.

Satya Almasian, Dennis Aumiller, and Michael Gertz. 2021. Bert got a date: Introducing transformers to temporal tagging. *arXiv preprint arXiv:2109.14927*.

Satya Almasian, Dennis Aumiller, and Michael Gertz. 2022. Time for some german? pre-training a transformer-based temporal tagger for german. *Text2Story@ ECIR*, 3117.

Masayuki Asahara, Sachi Yasuda, Hikari Konishi, Mizuho Imada, and Kikuo Maekawa. 2013. BCCWJ-TimeBank: Temporal and event information annotation on Japanese text. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, pages 206–214, Taipei, Taiwan. Department of English, National Chengchi University.

Miguel Ballesteros, Rishita Anubhai, Shuai Wang, Nima Pourdamghani, Yogarshi Vyas, Jie Ma, Parminder Bhatia, Kathleen McKeown, and Yaser Al-Onaizan. 2020. Severing the edge between before and after: Neural architectures for temporal ordering of events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5412–5417, Online. Association for Computational Linguistics.

Harsimran Bedi, Sangameshwar Patil, and Girish Palshikar. 2021. Temporal question generation from history text. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 408–413, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, Denver, Colorado. Association for Computational Linguistics.

Steven Bethard and Jonathan Parker. 2016. A semantically compositional annotation scheme for time normalization. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3779–3786, Portorož, Slovenia. European Language Resources Association (ELRA).

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

André Bittar, Pascal Amsili, Pascal Denis, and Laurence Danlos. 2011. French TimeBank: An ISO-TimeML annotated reference corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 130–134, Portland, Oregon, USA. Association for Computational Linguistics.

Alice Bracchi, Tommaso Caselli, and Irina Prodanof. 2016. Enrichring the ita-timebank with narrative containers. In *Proceedings of Third Italian Conference on Computational Linguistics CLiC-it 2016*, pages 83–88. Accademia University Press.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, and Wei Bi. 2021. Uncertainty-aware self-training for semi-supervised event temporal relation extraction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2900–2904.

Yuwei Cao, William Groves, Tanay Kumar Saha, Joel Tetreault, Alejandro Jaimes, Hao Peng, and Philip Yu. 2022. XLTime: A cross-lingual knowledge transfer framework for temporal expression extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1931–1942, Seattle, United States. Association for Computational Linguistics.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Angel X. Chang and Christopher Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3735–3740, Istanbul, Turkey. European Language Resources Association (ELRA).

Sanxing Chen, Guoxin Wang, and Börje Karlsson. 2019. Exploring word representations on time expression recognition. *Microsoft Research Asia, Tech. Rep.*

Fei Cheng, Masayuki Asahara, Ichiro Kobayashi, and Sadao Kurohashi. 2020. Dynamically updating event representations for temporal relation classification with multi-category learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1352–1357, Online. Association for Computational Linguistics.

Fei Cheng and Yusuke Miyao. 2018. Inducing temporal relations from time anchor annotation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1833–1843, New Orleans, Louisiana. Association for Computational Linguistics.

Jeremy R. Cole, Aditi Chaudhary, Bhuwan Dhingra, and Partha Talukdar. 2023. Salient span masking for temporal understanding. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3052–3060, Dubrovnik, Croatia. Association for Computational Linguistics.

Francisco Costa and António Branco. 2012. TimeBankPT: A TimeML annotated corpus of Portuguese. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3727–3734, Istanbul, Turkey. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sara Di Bartolomeo, Aditeya Pandey, Aristotelis Leventidis, David Saffo, Uzma Haque Syeda, Elin Carstensdottir, Magy Seif El-Nasr, Michelle A Borkin, and Cody Dunne. 2020. Evaluating the effect of timeline shape on visualization task performance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Dmitriy Dligach, Steven Bethard, Timothy Miller, and Guergana Savova. 2022. Exploring text representations for generative temporal relation extraction. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 109–113, Seattle, WA. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Alon Eirew, Kfir Bar, and Ido Dagan. 2025. Beyond pairwise: Global zero-shot temporal graph generation. *arXiv preprint arXiv:2502.11114*.

Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019a. Deep structured neural network for event temporal relation extraction. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 666–106, Hong Kong, China. Association for Computational Linguistics.

Rujun Han, Mengyue Liang, Bashar Alhafni, and Nanyun Peng. 2019b. Contextualized word embeddings enhanced event temporal relation extraction for story understanding. *arXiv preprint arXiv:1904.11942*.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019c. Joint event and temporal relation extraction with shared representations and structured prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 434–444, Hong Kong, China. Association for Computational Linguistics.

Rujun Han, Xiang Ren, and Nanyun Peng. 2021. ECONET: Effective continual pretraining of language models for event temporal reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5367–5380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rujun Han, Yichao Zhou, and Nanyun Peng. 2020. Domain knowledge empowered structured neural net for end-to-end event temporal relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5717–5729, Online. Association for Computational Linguistics.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Jiafeng Guo, and Xueqi Cheng. 2025. Large language model-based event relation extraction with rationales. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7484–7496.

Quzhe Huang, Yutong Hu, Shengqi Zhu, Yansong Feng, Chang Liu, and Dongyan Zhao. 2023. More than classification: A unified framework for event temporal relation extraction. *arXiv preprint arXiv:2305.17607*.

EunJeong Hwang, Jay-Yoon Lee, Tianyi Yang, Dhruvesh Patel, Dongxu Zhang, and Andrew McCallum. 2022. Event-event relation extraction using probabilistic box embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 235–244, Dublin, Ireland. Association for Computational Linguistics.

Lis Kanashiro Pereira. 2022. Attention-focused adversarial training for robust temporal reasoning. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7352–7359, Marseille, France. European Language Resources Association.

Allen Kim, Charuta Pethe, and Steve Skiena. 2020. What time is it? temporal analysis of novels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9076–9086, Online. Association for Computational Linguistics.

Lukas Lange, Jannik Strötgen, Heike Adel, and Dietrich Klakow. 2023. Multilingual normalization of temporal expressions with masked language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1174–1186, Dubrovnik, Croatia. Association for Computational Linguistics.

Egoitz Laparra, Xin Su, Yiyun Zhao, Özlem Uzuner, Timothy Miller, and Steven Bethard. 2021. SemEval-2021 task 10: Source-free domain adaptation for semantic processing. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 348–356, Online. Association for Computational Linguistics.

Egoitz Laparra, Dongfang Xu, Ahmed Elsayed, Steven Bethard, and Martha Palmer. 2018. SemEval 2018 task 6: Parsing time normalizations. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 88–96, New Orleans, Louisiana. Association for Computational Linguistics.

Artuur Leeuwenberg and Marie-Francine Moens. 2019. A survey on temporal reasoning for temporal information extraction from text. *Journal of Artificial Intelligence Research*, 66:341–380.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Manling Li, Tengfei Ma, Mo Yu, Lingfei Wu, Tian Gao, Heng Ji, and Kathleen McKeown. 2021. Timeline summarization based on event graph compression via time-aware optimal transport. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6443–6456, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, et al. 2022. A survey on deep learning event extraction: Approaches and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Chae-Gyun Lim, Young-Seob Jeong, and Ho-Jin Choi. 2018. Korean TimeBank including relative temporal information. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Chae-Gyun Lim, Young-Seob Jeong, and Ho-Jin Choi. 2019. Survey of temporal information extraction. *Journal of Information Processing Systems*, 15(4):931–956.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Farig Sadeque, Steven Bethard, and Guergana Savova. 2020. A BERT-based one-pass multi-task model for clinical temporal relation extraction. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 70–75, Online. Association for Computational Linguistics.

Jian Liu, Jinan Xu, Yufeng Chen, and Yujie Zhang. 2021. Discourse-level event temporal ordering with uncertainty-guided graph completion. In *IJCAI*, pages 3871–3877.

Xiaochen Liu and Yanan Zhang. 2025. Etimeline: An extensive timeline generation dataset based on large language model. *arXiv preprint arXiv:2502.07474*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hieu Man, Nghia Trung Ngo, Linh Ngo Van, and Thien Huu Nguyen. 2022. Selecting optimal context sentences for event-event relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11058–11066.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad Morariu, Quan Hung Tran, and Dinesh Manocha. 2021. TIMERS: Document-level temporal relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 524–533, Online. Association for Computational Linguistics.

Puneet Mathur, Vlad Morariu, Verena Kaynig-Fittkau, Jiuxiang Gu, Franck Dernoncourt, Quan Tran, Ani Nenkova, Dinesh Manocha, and Rajiv Jain. 2022. DocTime: A document-level temporal dependency graph parser. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 993–1009, Seattle, United States. Association for Computational Linguistics.

Ulrike May, Karolina Zaczynska, Julián Moreno-Schneider, and Georg Rehm. 2021. Extraction and normalization of vague time expressions in German. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 114–126, Düsseldorf, Germany. KONVENS 2021 Organizers.

Pawel Mazur and Robert Dale. 2010. WikiWars: A new corpus for research on temporal expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 913–922, Cambridge, MA. Association for Computational Linguistics.

Anne-Lyse Minard, Manuela Speranza, Ruben Urizar, Begoña Altuna, Marieke van Erp, Anneleen Schoen, and Chantal van Son. 2016. MEANTIME, the NewsReader multilingual event and time corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4417–4422, Portorož, Slovenia. European Language Resources Association (ELRA).

Sajad Mirzababaei, Amir Hossein Kargaran, Hinrich Schütze, and Ehsaneddin Asgari. 2022. Hengam: An adversarially trained transformer for Persian temporal tagging. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1013–1024, Online only. Association for Computational Linguistics.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics.

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. TDDiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249, Stockholm, Sweden. Association for Computational Linguistics.

Marta Guerrero Nieto, Roser Saurí, and Miguel Angel Bernabé Poveda. 2011. Modes timebank: A modern spanish timebank corpus. *Procesamiento del lenguaje natural*, 47:259–267.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6203–6209, Hong Kong, China. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Wanting Ning, Lishuang Li, Xueyang Qin, Yubo Feng, and Jingyao Tang. 2024. Temporal cognitive tree: A hierarchical modeling approach for event temporal relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 855–864, Miami, Florida, USA. Association for Computational Linguistics.

Jingcheng Niu, Saifei Liao, Victoria Ng, Simon De Montigny, and Gerald Penn. 2024. ConTempo: A unified temporally contrastive framework for temporal relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1521–1533, Bangkok, Thailand. Association for Computational Linguistics.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Amy L Olex and Bridget T McInnes. 2021. Review of temporal reasoning in the clinical domain for timeline extraction: Where we are and where we need to be. *Journal of biomedical informatics*, 118:103784.

Amy L Olex and Bridget T McInnes. 2022. Temporal disambiguation of relative temporal expressions in clinical texts. *Frontiers in Research Metrics and Analytics*, 7:1001266.

Lis Pereira, Fei Cheng, Masayuki Asahara, and Ichiro Kobayashi. 2021. ALICE++: Adversarial training for robust and effective temporal reasoning. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 373–382, Shanghai, China. Association for Computational Lingustics.

James Pustejovsky. 2003. Timeml: Robust specification of event and temporal expressions in text. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5), 2003*.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Anna Rogers, Marzena Karpinska, Ankita Gupta, Vladislav Lialin, Gregory Smelkov, and Anna Rumshisky. 2019. Narrativetime: Dense temporal annotation on a timeline. *arXiv preprint arXiv:1908.11443*.

Hayley Ross, Jonathon Cai, and Bonan Min. 2020. Exploring Contextualized Neural Language Models for Temporal Dependency Parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8548–8553, Online. Association for Computational Linguistics.

Vered Shwartz. 2022. Good night at 4 pm?! time expressions in different cultures. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2842–2853, Dublin, Ireland. Association for Computational Linguistics.

Jannik Strötgen, Thomas Bögel, Julian Zell, Ayser Armiti, Tran Van Canh, and Michael Gertz. 2014. Extending HeidelTime for temporal expressions referring to historic dates. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2390–2397, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, Uppsala, Sweden. Association for Computational Linguistics.

Jannik Strötgen, Anne-Lyse Minard, Lukas Lange, Manuela Speranza, and Bernardo Magnini. 2018. KRAUTS: A German temporally annotated news corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Xin Su, Phillip Howard, Nagib Hakim, and Steven Bethard. 2023. Fusing temporal graphs into transformers for time-sensitive question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 948–966, Singapore. Association for Computational Linguistics.

Xin Su, Yiyun Zhao, and Steven Bethard. 2021. The University of Arizona at SemEval-2021 task 10: Applying self-training, active learning and data augmentation to source-free domain adaptation. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 458–466, Online. Association for Computational Linguistics.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2021. Extracting event temporal relations via hyperbolic geometry. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8065–8077, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2023. Event temporal relation extraction with Bayesian translational model. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1125–1138, Dubrovnik, Croatia. Association for Computational Linguistics.

Sun Tiesen and Li Lishuang. 2022. Improving event temporal relation classification via auxiliary label-aware contrastive learning. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 861–871, Nanchang, China. Chinese Information Processing Society of China.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. 2007. SemEval-2007 task 15: TempEval temporal relation identification. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 75–80, Prague, Czech Republic. Association for Computational Linguistics.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden. Association for Computational Linguistics.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706, Online. Association for Computational Linguistics.

Haoyu Wang, Hongming Zhang, Yuqian Deng, Jacob Gardner, Dan Roth, and Muhao Chen. 2023. Extracting or guessing? improving faithfulness of event temporal relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 541–553, Dubrovnik, Croatia. Association for Computational Linguistics.

Liang Wang, Peifeng Li, and Sheng Xu. 2022. DCT-centered temporal relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2087–2097, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Haoyang Wen and Heng Ji. 2021. Utilizing relative event time to enhance event-event temporal relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10431–10437, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853*.

Sen Yang, Xin Li, Lidong Bing, and Wai Lam. 2023. Once upon a *time* in *graph*: Relative-time pretraining for complex temporal reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11879–11895, Singapore. Association for Computational Linguistics.

Hao-Ren Yao, Luke Breitfeller, Aakanksha Naik, Chunxiao Zhou, and Carolyn Rose. 2024a. Distilling multi-scale knowledge for event temporal relation extraction. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2971–2980.

Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli Goldner, and Guergana Savova. 2024b. Overview of the 2024 shared task on chemotherapy treatment timeline extraction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 557–569, Mexico City, Mexico. Association for Computational Linguistics.

Jiarui Yao, Haoling Qiu, Bonan Min, and Nianwen Xue. 2020. Annotating Temporal Dependency Graphs via Crowdsourcing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5368–5380, Online. Association for Computational Linguistics.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

Alessandra Zarcone, Touhidul Alam, and Zahra Kolagar. 2020. PATE: A corpus of temporal expressions for the in-car voice assistant domain. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 523–530, Marseille, France. European Language Resources Association.

Shuaicheng Zhang, Qiang Ning, and Lifu Huang. 2022. Extracting temporal event relation with syntax-guided graph transformer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 379–390, Seattle, United States. Association for Computational Linguistics.

Yuchen Zhang and Nianwen Xue. 2018. Structured interpretation of temporal relations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yuchen Zhang and Nianwen Xue. 2019. Acquiring structured temporal representation via crowdsourcing: A feasibility study. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 178–185, Minneapolis,

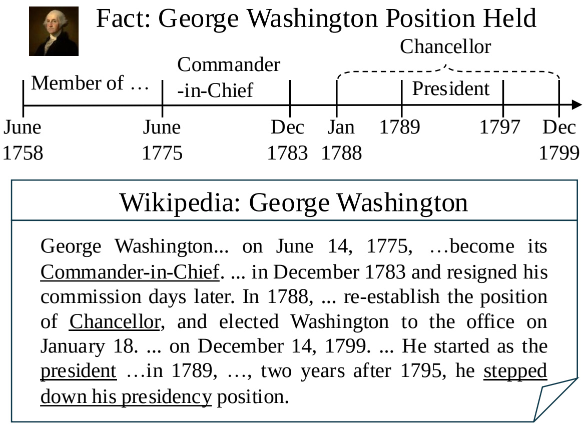

Figure 2: A snippet from George Washington's Wikipedia page and a timeline regarding his positions.

Minnesota. Association for Computational Linguistics.

Xinyu Zhao, Shih-Ting Lin, and Greg Durrett. 2021. Effective distant supervision for temporal relation extraction. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 195–203, Kyiv, Ukraine. Association for Computational Linguistics.

Jie Zhou, Shenpo Dong, Hongkui Tu, Xiaodong Wang, and Yong Dou. 2022. RSGT: Relational structure guided temporal relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2001–2010, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yichao Zhou, Yu Yan, Rujun Han, J Harry Caufield, Kai-Wei Chang, Yizhou Sun, Peipei Ping, and Wei Wang. 2021. Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14647–14655.

A Timeline Examples

We present in Figure 2 a snippet from George Washington's Wikipedia page alongside the corresponding timeline of his position changes.

B Evaluation Metrics

In temporal IE, the evaluation method from TEMPEVAL-3 (UzZaman et al., 2013) is the most widely adopted standard. This evaluation method calculates the standard precision (P), recall (R), and F1 score (F) between the system predictions (System) and the gold annotations (Reference) as follows:

$$P = \frac{|\text{System} \cap \text{Reference}|}{|\text{System}|} \tag{1}$$

$$R = \frac{|\text{System} \cap \text{Reference}|}{|\text{Reference}|} \tag{2}$$

$$F = 2 \cdot \frac{P \cdot R}{P + R} \tag{3}$$

In time expression identification, "System" refers to the time expressions identified by the system, while "Reference" refers to the annotated gold time expressions. In time expression normalization, "System" and "Reference" refer to the system-normalized time expressions and the gold annotated normalized expressions, respectively. If calculating the end-to-end time expression normalization score, "System" only involves the correctly identified time expressions.

For the temporal relation extraction task, the TEMPEVAL-3 evaluation method calculates the temporal awareness scores. This is achieved by performing a graph closure operation on the gold temporal graph based on temporal transitivity rules (to incorporate all potential temporal relations) and reducing the predicted temporal relation graph (to remove duplicate relations). These steps are completed before calculating the standard scores. Here, "System" denotes the temporal relations predicted by the system, while "Reference" is the gold annotated temporal relations.

C Datasets Summary

We summarize the temporal IE datasets in Table 1. The first section is based on the most widely used TimeML annotation framework, while the second section covers those that adopt all other annotation frameworks.

D Timeline-focused Datasets

A notable trend in temporal IE dataset development is the emergence of timeline-focused annotation frameworks that offer more comprehensive and coherent temporal representations compared to traditional approaches. For timeline-centric annotation, Rogers et al. (2019) propose NarrativeTIME, which enables dense, full-coverage temporal relation annotation. Unlike the pairwise TLINK annotation in TimeML, NarrativeTIME constructs coherent narrative timelines, supports underspecification via event types and timeline branches, and achieves significantly higher annotation density. Similarly, Liu and Zhang (2025) introduce ETimeline, a large-scale bilingual (English/Chinese) timeline dataset comprising over 600 timelines and 13,878 annotated event entries, spanning diverse domains from

Name	Framework	Domain	Lang	Tasks
		TimeML-Based		
TimeBank (Pustejovsky, 2003)	TimeML	Newswire	EN	I, N, R
TempEval-1 (Verhagen et al., 2007)	TimeML	Newswire	EN	I, N, R
TempEval-2 (Verhagen et al., 2010)	TimeML	Newswire	ZH, EN, IT, FR, KR, ES	I, N, R
Spanish TimeBank (Nieto et al., 2011)	TimeML	Historiography	ES	I, N
French TimeBank (Bittar et al., 2011)	ISO-TimeML	Newswire	FR	I, N, R
Portuguese TimeBank (Costa and Branco, 2012)	TimeML	Newswire	PT	I, N, R
i2b2-2012 (Sun et al., 2013)	Thyme-TimeML	Clinical	EN	I, N, R
TempEval-3 (UzZaman et al., 2013)	TimeML	Newswire	EN, ES	I, N, R
TimeBank-Dense (Chambers et al., 2014)	TimeML	Newswire	EN	I, N, R
Japanese TimeBank (Asahara et al., 2013)	ISO-TimeML	Publication, Library, Special purpose	JA	I, N, R
AncientTimes (Strötgen et al., 2014)	TimeML	Wikipedia	EN, DE, NL, ES, FR, IT, AR, VI	I, N
THYME-2015 (Bethard et al., 2015)	Thyme-TimeML	Clinical	EN	I, N, R
THYME-2016 (Bethard et al., 2016)	Thyme-TimeML	Clinical	EN	I, N, R
Richer Event Description (O'Gorman et al., 2016)	Thyme-TimeML	Newswire, Forum Discussions	EN	I, N, R
Italian TimeBank (Bracchi et al., 2016)	TimeML	Newswire	IT	I, N, R
MeanTime (Minard et al., 2016)	ISO-TimeML	Newswire	EN, IT, ES, NL	I, N, R
THYME-2017 (Bethard et al., 2017)	Thyme-TimeML	Clinical	EN	I, N, R
Event StoryLine (Caselli and Vossen, 2017)	TimeML	Story	EN	I, N, R
MATRES (Ning et al., 2018)	TimeML	Newswire	EN	I, R
Korean TimeBank (Lim et al., 2018)	TimeML	Wikipedia	KR	I, N, R
German Temporal Expression (Strötgen et al., 2018)	TimeML	Newswire	DE	I, N
TDDiscourse (Naik et al., 2019)	TimeML	Newswire	EN	R
PATE (Zarcone et al., 2020)	TimeML	Voice Assistant	EN	I, N
German VTEs (May et al., 2021)	ISO-TimeML	Newswire	DE	I, N
		Other Annotation Framework-based		
WikiWars (Mazur and Dale, 2010)	TIMEX2	Wikipedia	EN, DE	I, N
SCATE (Bethard and Parker, 2016; Laparra et al., 2018)	SCATE	Newswire, Clinical	EN	I, N
CaTeRS (Mostafazadeh et al., 2016)	CaTeRS	Commonsense Stories	EN	R
TORDER (Cheng and Miyao, 2018)	TORDER	Newswire	EN	R
Temporal Dependency Tree (Zhang and Xue, 2018, 2019)	Temporal Dependency Tree	Newswire, Narratives	ZH	R
Temporal Dependency Graph (Yao et al., 2020)	Temporal Dependency Graph	Newswire	EN	R

Table 1: Overview of datasets and their schemas, domains, languages (EN: English, DE: German, NL: Dutch, ES: Spanish, FR: French, IT: Italian, AR: Arabic, VI: Vietnamese, JA: Japanese, PT: Portuguese, ZH: Chinese, KR: Korean), and tasks (I: identification, N: time expression normalization, R: temporal relation extraction).

March 2020 to April 2024. Created using an LLM-assisted annotation approach, ETimeline represents a significant resource for cross-lingual timeline construction and temporal reasoning across news domains.

E Temporal Relation Extraction Methods Summary

We summarize the temporal relation extraction methods we review in Table 2.

F Discussion on Future Directions

In the previous sections, we have identified the following research opportunities in the field of temporal IE:

- Enrich annotation frameworks (Section 3.3), e.g., representing event arguments or expanding formal semantic systems like SCATE.
- Improve dataset diversity (Section 3.3), e.g., annotating more domains beyond newswire.
- Explore generative approaches (Sections 4.2 and 5.3), e.g., new input-output formulations, new fine-tuning strategies.
- Develop public tools and benchmarks (Sec-

Work	Approach	Base Model	Evaluation Datasets	Knwl	Rbst	Avl
Lin et al. (2019)	Discr.	BERT	THYME	✗	✓	✗
Han et al. (2019a)	Discr.	BERT	TimeBank-Dense, MATRES	✓	✗	✗
Ning et al. (2019)	Discr.	BERT	TimeBank-Dense, MATRES	✓	✗	✗
Han et al. (2019c)	Discr.	BERT	TimeBank-Dense, MATRES	✓	✓	✗
Han et al. (2019b)	Discr.	BERT	Richer Event Description, CaTeRS	✓	✓	✗
Lin et al. (2020)	Discr.	BERT	THYME	✗	✓	✗
Cheng et al. (2020) (SEC)	Discr.	BERT	Japanese-Timebank, TimeBank-Dense	✓	✓	✗
Ross et al. (2020)	Discr.	BERT	Temporal Dependency Tree	✓	✗	✗
Ballesteros et al. (2020)	Discr.	RoBERTa	MATRES	✗	✓	✗
Han et al. (2020)	Discr.	RoBERTa	i2b2-2012, TimeBank-Dense	✓	✓	✗
Wang et al. (2020)	Discr.	RoBERTa	MATRES	✓	✗	✗
Zhao et al. (2021)	Discr.	RoBERTa	MATRES	✗	✓	✓
Zhou et al. (2021) (CTRL-PG)	Discr.	BERT	i2b2-2012, TimeBank-Dense	✓	✗	✗
Cao et al. (2021) (UAST)	Discr.	RoBERTa	MATRES, TimeBank-Dense	✗	✓	✗
Tan et al. (2021)	Discr.	RoBERTa	MATRES	✓	✗	✗
Mathur et al. (2021) (TIMERS)	Discr.	BERT	TimeBank-Dense, MATRES, TDDiscourse	✓	✗	✗
Liu et al. (2021)	Discr.	BERT	TimeBank-Dense, TDDiscourse	✓	✗	✗
Wen and Ji (2021)	Discr.	RoBERTa	MATRES	✓	✗	✗
Pereira et al. (2021) (ALICE++)	Discr.	RoBERTa	MATRES, TimeML	✗	✓	✗
Han et al. (2021) (ECONET)	Discr.	RoBERTa/BERT	TimeBank-Dense, MATRES, Richer Event Description	✗	✓	✓
Kanashiro Pereira (2022) (ML-ALICE)	Discr.	RoBERTa	MATRES, TimeML	✗	✓	✗
Wang et al. (2022) (DTRE)	Discr.	RoBERTa	TimeBank-Dense, TDDiscourse	✓	✓	✗
Mathur et al. (2022) (DocTime)	Discr.	BERT	Temporal Dependency Tree	✓	✓	✗
Hwang et al. (2022) (BERE)	Discr.	RoBERTa	MATRES, Event StoryLine	✓	✗	✗
Dligach et al. (2022)	Gen	BART/T5	THYME	✗	✗	✗
Wang et al. (2023)	Discr.	BigBird	MATRES, TDDiscourse	✓	✓	✗
Zhang et al. (2022)	Discr.	BERT	MATRES, TimeBank-Dense	✓	✗	✗
Tiesen and Lishuang (2022) (TempACL)	Discr.	BERT	TimeBank-Dense, MATRES	✗	✓	✗
Zhou et al. (2022) (RSGT)	Discr.	RoBERTa	TimeBank-Dense, MATRES	✓	✗	✗
Man et al. (2022) (SCS-EERE)	Discr.	RoBERTa	MATRES, TDDiscourse	✓	✗	✗
Yuan et al. (2023)	Gen	ChatGPT	TimeBank-Dense, MATRES, TDDiscourse	✗	✗	✗
Huang et al. (2023)	Discr.	BERT/RoBERTa	TimeBank-Dense, MATRES	✓	✗	✗
Tan et al. (2023) (Bayesian-Trans)	Discr.	BART	MATRES, imeBank-Dense	✓	✗	✓
Niu et al. (2024) (ConTempo)	Discr.	RoBERTa	TimeBank-Dense, MATRES	✓	✓	✗

Table 2: Overview of research on temporal relation extraction. “Knwl” represents the inclusion of external knowledge. “Rbst" refers to the application of methods to enhance model robustness. “Avl” indicates whether the model is publicly available. Symbols ✓ and ✗ indicate the presence or absence of a feature, respectively.

tions 4.2 and 5.3), e.g., publish temporal IE models and datasets to the public repositories

- Explore new applications (Section 6.2), e.g., the utility of extracted timelines when visualized for human-computer interaction.

F.1 Enrich Annotation Frameworks and Improve the Domain Diversity of Datasets

Current annotation frameworks, such as TimeML, often produce temporal graphs composed of temporal relations and temporal entities, as illustrated in Figure 1. However, these temporal graphs are challenging to interpret independently or use directly for temporal reasoning without extensive context. One future direction could be to integrate richer content into end-to-end temporal IE annotation frameworks. One example is incorporating entity relation extraction and full event extraction (including triggers and arguments) from the general domain to construct a more complete temporal graph. This concept has begun to emerge in the literature, as seen in Li et al. (2021). Yet, that work mainly integrates existing temporal IE tools with general domain IE tools without proposing a well-defined annotation framework. Another example is to develop user-friendly frameworks like SCATE, which, unlike TimeML, outputs temporal intervals that can be directly mapped onto a timeline given a temporal expression. However, SCATE primarily focuses on the normalization of time expressions. Expanding its scope to include the normalization of a broader range of temporal content, such as events and sentences, could significantly widen its applicability.

Furthermore, future efforts could focus on expanding the domains covered by existing datasets to mitigate the domain bias present in current datasets. For example, the Thyme datasets represent an adaptation of TimeML to better suit the medical field's representation of temporal relations between events and times. Yet, such efforts to adapt and improve annotation frameworks for additional fields are still scarce. Therefore, adapting existing annotation frameworks to a broader range of domains to enhance the domain diversity of datasets represents a potential future research direction.

F.2 Improve the Application of Generative LLMs

The application of generative LLMs in the field of time expression identification, normalization, and temporal relation extraction remains underexplored. Given the proven capabilities of LLMs like ChatGPT and LLAMA3 across various tasks, it is logical to probe their potential within the realm of temporal IE. Whether it involves leveraging new prompting methods or fine-tuning strategies for specific tasks, there is ample room for innovation.

However, it is important to emphasize that while these models excel in generating unstructured text when applied to temporal IE, it is imperative to specially design suitable input-output formats. Such designs are intended to enable generative LLMs, which are typically used for producing unstructured text, to also effectively output structured temporal information.

F.3 Develop Public Toolkits and Evaluation Benchmarks

We believe that one key reason Transformer-based temporal IE models have not been widely adopted might be the absence of a publicly available code repository that facilitates easier access to models and data. For example, HuggingFace [1] provides language model heads or pipelines suitable for various tasks, allowing users to easily download and deploy trained models on any dataset directly from the HuggingFace Hub. A future research direction should involve establishing such a repository or pushing models/datasets to HuggingFace Hub for the temporal IE tasks to enhance the reproducibility and applicability of research. Another important direction is to create a public and test-set concealed benchmark for a more equitable comparison of existing work. In most existing works, although metrics such as F1 scores, precision, and recall are commonly computed, the specific implementations can vary. For instance, in Kanashiro Pereira (2022), only the "before" and "after" relationships are evaluated for relation extraction performance, whereas Zhang et al. (2022) includes all temporal relationships except "vague" in their evaluation.

F.4 Explore More Application Directions

In reviewing the application of temporal IE systems, we observe that current research primarily focuses on aiding "models" in temporal reasoning to enhance their performance in other tasks. Future research in temporal IE should not only continue to support model performance improvement but should also pay more attention to serving humans and enhancing its practical value. A promising application direction is visualizing timelines in human-computer interaction (HCI) scenarios. The visualization results of existing temporal graphs are often challenging for human users to interpret. For instance, visualizing the temporal graph of any document in the TimeBank-Dense dataset might result in a graph densely populated with points and lines, offering little help for users to comprehend the progression of events within the text.

User studies, such as those conducted by Di Bartolomeo et al. (2020), have revealed the importance of visualization forms of timelines for user understanding. Consequently, temporal IE research should also consider incorporating user research on temporal graphs to guide the design of temporal IE methods, such as how to represent standardized time expressions, identify which types of temporal relations most effectively facilitate time understanding, and determine the best ways to present this information. By addressing these problems, the extraction and representation of temporal in-

[1] https://huggingface.co/

formation can be more closely aligned with user needs, enhancing its application value in HCI.

G Comparison with Previous Surveys

Our survey offers several key advancements over previous reviews in the field of temporal information extraction. Prior surveys such as Lim et al. (2019) and Leeuwenberg and Moens (2019) provide only brief mentions of standard datasets like TimeBank and TempEval, and largely predate the Transformer era. More recent reviews in the clinical domain—such as Alfattni et al. (2020) and Olex and McInnes (2021)—present more detailed dataset descriptions but are limited to clinical texts and do not cover resources from other domains.

In contrast, our survey compiles and categorizes 32 datasets across multiple domains (newswire, clinical, Wikipedia, narratives) and 15 languages, structured by annotation framework (TimeML-based vs. alternative schemas such as SCATE, temporal dependency trees, or CaTeRS). We provide a systematic analysis of dataset diversity, domain bias, language coverage, and annotation schema. Notably, we quantitatively analyze dataset bias, identifying that 63% of current datasets come from the newswire domain, and highlight underexplored areas such as the low representation of historical and non-news domains.

Our work specifically focuses on the Transformer era, providing in-depth analysis of how these architectures are applied to temporal IE tasks, examination of fine-tuning strategies, and discussion of how pre-trained language models capture temporal information. We also offer a broader scope in terms of domain and language coverage compared to previous works that focus on specific domains or primarily discuss English-language resources.

This broader treatment of datasets and methods is intentional. Since Transformer-based approaches often depend heavily on annotated corpora for fine-tuning or benchmarking, a full understanding of available datasets and their annotation assumptions is crucial to contextualizing methodological advances in temporal information extraction.

H Comparative Analysis of Temporal Relation Extraction Methods

This appendix provides a detailed comparative analysis of different methodological approaches in temporal relation extraction, examining their strengths, limitations, and trade-offs.

H.1 Methodological Approaches Comparison

Table 3 presents a systematic comparison of major methodological categories in temporal relation extraction.

Table 4 presents a more detailed comparison between discriminative and generative methods. The consistent underperformance of generative approaches suggests the field has not yet found optimal ways to leverage LLMs for temporal relation extraction. Current evidence (Yuan et al., 2023; Huang et al., 2023) shows that even with advanced prompting strategies, LLMs achieve substantially lower F1 scores compared to fine-tuned BERT-based models. The trade-off currently favors discriminative models for accuracy-critical applications, while generative approaches may be preferred when flexibility, explainability, or few-shot learning are priorities.

H.2 Dataset Scale Analysis

A critical limitation in temporal IE research is the constrained scale of available datasets. The three most frequently used datasets for temporal relation extraction contain:

- MATRES: 274 documents
- TimeBank-Dense: 36 documents
- TDDiscourse: 34 documents
- Total: 344 documents

This scale is one to two orders of magnitude smaller than comparable IE datasets in related NLP tasks, which typically contain 1,000-5,000 documents.

This limited scale has several implications:

1. Statistical Reliability: With only 36 documents in TimeBank-Dense, individual documents represent nearly 3% of the dataset, making performance metrics highly sensitive to individual annotations.
2. Overfitting Risk: Extensive hyperparameter tuning on such small datasets may lead to learning dataset-specific patterns rather than generalizable temporal reasoning.
3. Limited Diversity: Combined with the 63% newswire domain concentration (as documented in Section 3.3), the small scale severely limits assessment of model robustness.

Method Category	Representative Works	Strengths	Limitations
Commonsense Knowledge Integration	Ning et al. (2019), Wang et al. (2020), Tan et al. (2023)	• Captures human-intuitive event sequences • Improves implicit temporal reasoning • Better performance on narrative texts	• Requires external knowledge bases • Incomplete knowledge coverage • Domain-specific knowledge gaps
Syntactic/Semantic Knowledge	Wang et al. (2022), Mathur et al. (2021), Zhou et al. (2022)	• Leverages document structure • Captures long-range dependencies • Improves prediction coherence	• Depends on external parsing tools • Error propagation from parsing • Additional computational overhead
Temporal Rule Constraints	Hwang et al. (2022), Wang et al. (2020), Han et al. (2019a)	• Ensures logical consistency • Reduces impossible predictions • Global coherence improvement	• Too rigid for ambiguous cases • Difficulty handling exceptions • Complex inference procedures
Robustness Enhancement	Cao et al. (2021), Zhao et al. (2021), Pereira et al. (2021)	• Better cross-domain transfer • Reduced overfitting • More stable performance	• Increased training complexity • Additional data requirements • May sacrifice peak accuracy
Generative Approaches	Dligach et al. (2022), Yuan et al. (2023), Huang et al. (2023)	• Flexible output formats • Zero-shot capabilities • Leverages pre-trained LLMs	• Underperforms discriminative models • Requires careful prompt design • Output parsing challenges

Table 3: Comparative analysis of temporal relation extraction methodologies. Each category represents a distinct approach to addressing challenges in temporal IE.

Aspect	Discriminative Models	Generative Models
Performance	State-of-the-art on benchmarks	Consistently lower
Efficiency	Fast inference, millions of parameters	Slower inference, billions of parameters
Data Requirements	Requires substantial labeled data	Few-shot learning capabilities
Flexibility	Fixed relation types, requires retraining	Adaptable to new relations without retraining
Interpretability	Limited, attention weights only	Can provide natural language explanations

Table 4: Trade-offs between discriminative and generative approaches in temporal relation extraction.