TIMERS: Document-level Temporal Relation Extraction

Puneet Mathur
University of Maryland, College Park
puneetm@umd.edu

Rajiv Jain
Adobe Research
rajijain@adobe.com

Franck Dernoncourt
Adobe Research
dernonco@adobe.com

Vlad Morariu
Adobe Research
morariu@adobe.com

Quan Hung Tran
Adobe Research
qtran@adobe.com

Dinesh Manocha
University of Maryland, College Park
dmanocha@umd.edu

Abstract

We present **TIMERS** - a **TIME**, **R**hetorical and **S**yntactic-aware model for document-level temporal relation classification. Our proposed method leverages rhetorical discourse features and temporal arguments from semantic role labels, in addition to traditional local syntactic features, trained through a Gated Relational-GCN. Extensive experiments show that the proposed model outperforms previous methods by 5-18% on the TDDiscourse, TimeBank-Dense, and MATRES datasets due to our discourse-level modeling.

1 Introduction

Temporal relation extraction (TempRel) is a challenging task that involves determining the temporal order between two events in a text (Pustejovsky et al., 2003). Understanding the temporal ordering of events in a document plays a key role in downstream tasks such as timeline creation (Leeuwenberg and Moens, 2018), time-aware summarization (Noh et al., 2020), temporal question-answering (Ning et al., 2020), and temporal information extraction (Leeuwenberg and Moens, 2019).

Prior work focuses on extracting temporal relations between event pairs (a.k.a., *TLINKS*) present in the same sentence (*Intra-sentence TLINKS*) or adjacent sentences (*Inter-sentence TLINKS*), mostly ignoring document-level pairs (*Cross-document TLINKS*) (Reimers et al., 2016). Past works have used RNN (Cheng and Miyao, 2017; Meng et al., 2017; Goyal and Durrett, 2019; Ning et al., 2019; Han et al., 2019a,c,b, 2020b) and Transformer networks (Ballesteros et al., 2020; Zhao et al., 2020b) for encoding a few sentences or a short paragraph but do not capture long-range dependencies and multi-hop reasoning at the document-level. This shortcoming is shown in the TDDiscourse dataset (Naik et al., 2019), which was designed to highlight global discourse-level challenges, e.g., multi-hop chain reasoning, future or hypothetical events, and reasoning requiring world knowledge.

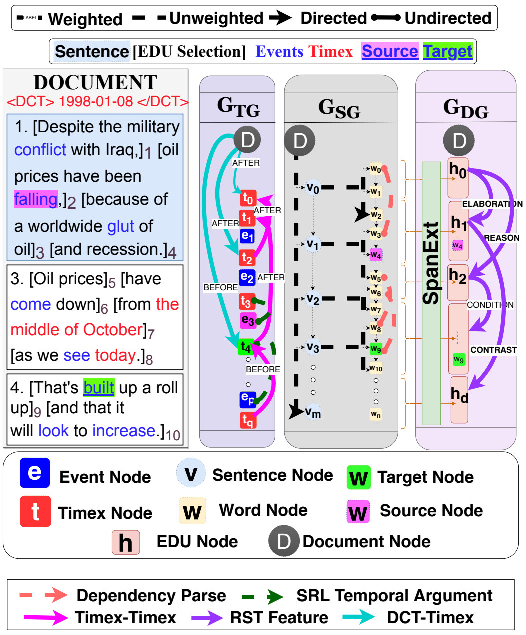
We propose **TIMERS** - a **TIME**, **R**hetorical, and **S**yntactic-aware model for document-level temporal relation extraction. TIMERS uses discourse features in the form of connections from Rhetorical Structure Theory (RST) parsers (Bhatia et al., 2015) to leverage long-range inter-sentential relationships. It also extends existing contextual embeddings with structural and syntactic dependency parse connections. Lastly, it uses timex-timex relations, *dct* (document creation date)-timex relations, and temporal arguments obtained via sentence-level semantic role labeling. These rhetorical, syntactic, and temporal features are learned through a modified version of Relational Graph Convolutional Networks (R-GCN) with a gating mechanism (GR-GCN) (Schlichtkrull et al., 2018), which learns highly relational data relationships in densely-connected graph networks.

Our **main contribution** is a document-level model that incorporates these three features to improve temporal relationship extraction. We obtain state-of-the-art performance across three datasets with **5-18% relative improvement**, showing improvement for events that require chain reasoning, causal prerequisite links, and future events.

2 Methodology

Let document D be defined as a sequence of n tokens $w_i \in W = \{w_1, \cdots, w_n\}$. The entire document is a list of m sentences $V = [v_1, \cdots, v_m]$. Each document has a set of p events $E = \{e_1, \cdots, e_p\}$ and q timexes $T = \{t_1, \cdots, t_q\}$, where $p, q \leq n$. The creation date of the document is represented by timestamp t_{DCT}. We denote the source and target events by e_s and e_t, respec-

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics
and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 524–533
August 1–6, 2021. ©2021 Association for Computational Linguistics



Weighted
Unweighted
Directed
Undirected
Sentence [EDU Selection] Events Timex Source Target
DOCUMENT
<DCT> 1998-01-08 </DCT>
1. [Despite the military conflict with Iraq,]1 [oil prices have been falling,]2 [because of a worldwide glut of oil]3 [and recession.]4
3. [Oil prices]5 [have come down]6 [from the middle of October]7 [as we see today.]8
4. [That's built up a roll up]9 [and that it will look to increase.]10
GTG
GSG
GDG
D
AFTER
BEFORE
SpanExt
ELABORATION
REASON
CONDITION
CONTRAST
Event Node
Sentence Node
Target Node
Timex Node
Word Node
Source Node
EDU Node
Document Node
Dependency Parse
SRL Temporal Argument
Timex-Timex
RST Feature
DCT-Timex

Figure 1: Three graphs are created from the input document. Time-aware Graph (G_{TG}): DCT-Timex associations, Timex-Timex associations, and Temporal Argument connections from semantic role labels; Syntactic-aware Graph (G_{SG}): structural and syntactic connections; and Rhetoric-aware Graph (G_{DG}): rhetorical relations between EDU's (h_i).

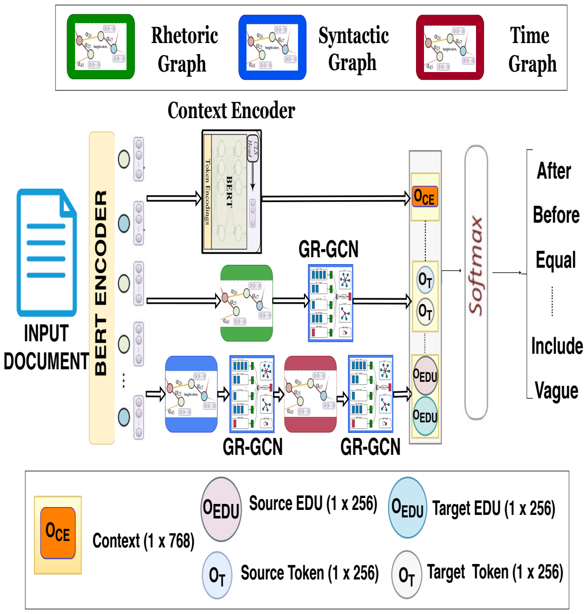


Figure 2: TIMERS learns rhetorical, syntactic, and temporal features through a Gated Relational-Graph Convolutional Networks (GR-GCN). The output of G_{SG} forms the input of G_{TG}. The output corresponding to the source and target nodes learned by G_{TG} (O_T) and G_{DG} (O_{EDU}) are concatenated with the output of the BERT based context encoder (O_{CE}), which forms the final output h_G that passes through the Softmax layer to predict the temporal relation.

tively. The task is to identify the temporal relation $y \in R$ between the source and target event in a multi-class classification setup, where R is the set of all possible temporal links (*TLINKs*).
To solve this task, our model (Fig.1) builds the **TIMERS**-graph, which consists of a Syntactic Graph (Sec.2.1), a Time Graph (Sec. 2.2), and a Rhetorical Graph (Sec.2.3). Each graph is learned through GR-GCN to extract the embeddings used for temporal relation extraction (Fig.2, Sec.2.4).

2.1 Syntactic-Aware Graph

The syntactic graph captures the document structure and word dependency. Our syntactic-aware graph ($\mathcal{G}_{SG}$) is made of separate nodes to represent the document D, each of its inherent sentences $v_i \in V$, and all the constituent words $w_i \in W$ of each sentence. The edges of the Syntactic Graph encode five relations: **(1) Document-Sentence Affiliation** and **(2) Sentence-Word Affiliation** model the hierarchical structure of the document through a directed edge from the document node to each sentence node and from a sentence node to each word in the sentence. **(3) Sentence-Sentence Adjacency** and **(4) Word-Word Adjacency** to preserve sequential ordering for consecutive sentence and word nodes. **(5) Word-Word Dependency** encodes the syntactical nature of the word-level relationships by adding an undirected edge between two word nodes if they share a parent-child relationship in the sentence-level dependency tree.

We use BERT to encode each w_i and obtain sentence embeddings $v_i^{'}$ by averaging the second-to-last hidden layer of BERT for each token. The document vector embedding $D_i^{'}$ was calculated as the average of all sentence embedding ($D_i^{'} = \sum_{i=0}^{m} v_i^{'}$).

2.2 Time-Aware Graph

When events are anchored to a specific time, it becomes easier to infer event relationships from their associated date and time. The time-aware graph ($\mathcal{G}_{TG}$) exploits this intuition and propagates relational information among events, timexes, and the Document Creation Time (*DCT*). The document node D is the node corresponding to the document creation date while the timexes t_i and events e_i are characterized by their corresponding word nodes in the Syntactic Graph. We design three types of edge connections: **(1) *DCT*-Timex Association:** exploit the ordering of timexes with respect to the document creation time through directed weighted edges from *DCT* to timexes. **(2) Timex-Timex Association:** capture inherent non-local timeline ordering between timex pairs by a directed weighted edge. **(3) Predicate-Temporal Argument:** anchor local temporal relations at the sentence level by connecting each event verb predicate to its temporal argument with a directed edge. The connections formed between temporal entities help navigate information from the source event to the target event while exploring interactions with other events, timexes, *dct*, and temporal arguments.

We calculate timestamps for timexes and the *DCT* from the annotated TimeML format of input documents. The weight of the *DCT*-timex and timex-timex edges is determined based on the temporal order of the entities $\{After, Before, Simultaneous, None\}$. We added *None* as a relation when one of the timestamps cannot be anchored in time.

2.3 Rhetorical-Aware Graph

We use discourse features based on Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) to leverage long-range inter-dependencies through a discourse tree. The rhetorical discourse tree of a document contains nodes of phrases, where each phrase (a.k.a, Elementary Discourse Unit or EDU) is contiguous, adjacent and non-overlapping. The interdependencies among EDUs are represented by conventional rhetorical relations (Mann, 1987), e.g. *Elaboration, Span, Condition, Attribution*. Prior work showed discourse features in the form of RST connections help leverage long-range document-level interactions between phrase units (Bhatia et al., 2015) and identify background-foreground events (Aldawsari et al., 2020).

Elementary Discourse Unit (EDU), a sub-sentence phrase unit, is the minimal selection unit for discourse segmentation of a document. We generate the document vector representations at EDU-level $h_i \in H = \{h_1, \cdots, h_d\}$ via the Self-Attentive Span Extractor (SpanExt) from Lee et al. (2017) over the BERT token embeddings. We use the converted dependency version of the tree to build the Rhetorical-aware graph ($\mathcal{G}_{DG}$) by treating every discourse dependency from the i-th EDU to the j-th EDU as a directed edge weighted by the type of the rhetorical relation.

2.4 Temporal Relation Extraction

Each graph is instantiated as a gated variant of Relational Graph Convolutional Networks (R-GCN) (Schlichtkrull et al., 2018), which we term as Gated Relational Graph Convolution Network (GR-GCN). GR-GCN propagates messages among the nodes to obtain a learned node representation and is inspired by (Zhang et al., 2020). Fig. 2 shows how the learned representations obtained from the syntactic-aware graph forms the input to the time-aware graph. For the time-aware graphs, the learned representations of nodes corresponding to the source event e_s and target event e_t are extracted (O_T). In the case of the rhetorical graphs, the span representations of the EDU span nodes corresponding to the source event (h_e) and target event (h_s) are extracted (O_{EDU}).

The output corresponding to the source and target nodes learnt by G_{TG} (O_T) and G_{DG} (O_{EDU}) are concatenated with output of BERT based context encoder (O_{CE}) (similar to BERT encoding in (Zhao et al., 2020a)): z_G = ReLU$(W[O_T; O_{EDU}; O_{CE}] + b)$. This is followed by a Softmax layer to predict temporal relations.

Dataset	Train	Validation	Test	Labels
TDDMan (Naik et al., 2019)	4000	650	1500	a, b, s, i, ii
TDDAuto (Naik et al., 2019)	32609	1435	4258	a, b, s, i, ii
MATRES (Ning et al., 2018a) ##	231	25	20	e,a,b,v
TimeBank-Dense (Cassidy et al., 2014)	4032	629	1427	a, b, s, i, ii, v

Table 1: Train/Val/Test data distribution for TDDMan, TDDAuto, MATRES, and TimeBank-Dense; a: After, b: Before, s: Simultaneous, i: Includes, ii: Is_included, v: Vague, e: Equal. (## Ning et al. (2019) use TimeBank and Aquaint for training, Platinum for test; 20% of train as validation)

Corpus	Model	F1
TB-Dense	Vashishtha et al. (2019)	56.6
	EventPlus (Ma et al., 2021)	64.5
	CTRL-PG (Zhou et al., 2020)	65.2
	DEER (Han et al., 2020a)	66.8
	TIMERS (ours)	**67.8**
MATRES	CogCompTime (Ning et al., 2018b)	66.6
	Goyal and Durrett (2019)	68.61
	BiLSTM+MAP (Han et al., 2019c)	75.5
	EventPlus (Ma et al., 2021)	75.5
	Wang et al. (2020)	78.8
	DEER (Han et al., 2020a)	79.3
	Zhao et al. (2020a)	79.6
	SMTL (Ballesteros et al., 2020)	81.6
	TIMERS (ours)	**82.3**

Table 2: Comparison of TIMERS with recent state-of-the-art models on TimeBank-Dense and MATRES dataset. TIMERS outperforms all recent top-performing systems.

3 Experiments

3.1 Data

We train and test our proposed model using the TDDMan and TDDAuto subsets of the TDDiscourse corpus (Naik et al., 2019), which was designed to explicitly focus on global discourse-level temporal ordering. We also train and evaluate our method on the MATRES and TimeBank-Dense datasets, both of which primarily consist of local TLINKs that occur in either the same or adjacent

	System	TDDMan			TDDAuto			MATRES			TB-Dense		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baselines	Majority	37.8	36.3	37.1	34.2	32.3	33.2	50.7	50.7	50.7	40.5	40.5	40.5
	CAEVO (Chambers et al., 2014)	32.3	10.7	16.1	61.1	32.6	42.5	-	-	-	49.9	46.6	48.2
	SP (Ning et al., 2017)	22.7	22.7	22.7	43.2	43.2	43.2	66.0	72.3	69.0	37.7	37.8	37.7
	SP+ILP (Ning et al., 2017)	23.9	23.8	23.8	46.4	45.9	46.1	71.3	82.1	76.3	58.4	58.4	58.4
	BiLSTM (Cheng and Miyao, 2017)	24.9	23.8	24.3	55.7	48.3	51.8	59.5	59.5	59.5	**63.9**	38.9	48.4
	BERT-base Transformer	36.5	37.1	37.5	62.0	61.7	62.3	65.6	78.1	77.2	59.7	60.7	62.2
	RoBERTa-base	35.7	36.5	37.1	60.6	62.7	61.6	77.3	79.0	78.9	58.1	57.6	61.9
	TIMERS (ours)	**43.7***	**46.7***	**45.5***	**64.3***	**72.7***	**71.1***	**81.1***	**84.6***	**82.3***	48.1	**65.2***	**67.8**
Ablation	TIMERS w\o Context Encoder	29.7	35.5	33.7	49.8	52.5	51.6	61.2	69.6	68.6	43.8	54.5	50.6
	TIMERS w\o $\mathcal{G}_{DG}$	39.6	39.6	41.8	61.7	66.8	65.4	71.8	79.1	79.7	51.4	63.0	63.3
	TIMERS w\o $\mathcal{G}_{SG}$	38.5	42.6	42.3	63.3	69.5	68.9	71.6	78.5	78.2	51.1	62.1	62.8
	TIMERS w\o $\mathcal{G}_{TG}$	37.5	39.8	39.5	58.7	68.3	67.1	72.8	78.5	77.7	50.5	62.9	61.8

Table 3: Results comparing performance of TIMERS with baselines and ablative components on TDDMan, TDDAuto, MATRES and TimeBank-Dense datasets. We adopt the BERT and RoBERTa implementation from (Ballesteros et al., 2020). * indicates statistical significance over BERT Transformer ($p \leq 0.005$) under Wilcoxon's Signed Rank test. Darker green represents better F1 performance on ablation studies. Bold denotes the best performing model. TIMERS improves substantially over all datasets. The ablation shows that context, discourse ($\mathcal{G}_{DG}$), and time-aware ($\mathcal{G}_{TG}$) graph encoders prove to be most beneficial.

sentences. Table 1 reports the data statistics and label distributions. (Naik et al., 2019) shows the distribution of the distance between event-pairs for all TLINKs in the TDD test set and explains that nearly 53% TLINKs in the TDD dataset comprise of event pairs that are more than 5 sentences apart. Like Cheng and Miyao (2017), we report results on non-vague labels of TimeBank-Dense. MATRES has no standard validation set. Hence, we follow the split used in (Ning et al., 2019).

3.2 Experimental Settings

Token Encoding:The word-level token representations are obtained by summing the corresponding BERT embeddings from the last 4 layers of pre-trained BERT-base encoder. **Syntactic Dependency Parser**: The dependency parse tree of individual sentences is obtained via SpaCy[1] to form word-word dependency connections in the syntactic-aware graph. **Semantic Role Labeller**: We extract semantic role labels using AllenNLP's SRL parser[2] that internally uses SRL-BERT (Shi and Lin, 2019) to obtain the temporal arguments corresponding to each verb event. **Timex Normalization**: Timex phrases are treated as a single unit for the purpose of graph construction by average pooling their BERT tokenized representations. Microsoft Recognizers-Text[3] is employed to normalize timexes and DCT date-time values. The normalized timex expressions are compared through Allen's interval algebra, where each timex has a start and an endpoint. The comparison is then made on the basis of the endpoints of the timexes, forming an edge going from earlier to later ending timex. **RST Discourse Parser**: We used the shift-reduce discourse parser proposed by Ji and Eisenstein (2014) to build the discourse tree [4], which is post-processed using *discoursegraphs* library[5] (Neumann, 2015) to build the rhetorical dependencies graph. Further implementation details can be found in the appendix.

3.3 Results

Table 3 compares our work to the baseline methods reported on the TDDMan, TDDAuto, MATRES, and TimeBank-Dense datasets. We also include results for BERT-based Transformer (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) following Ballesteros et al. (2020). To prevent truncation or memory errors otherwise caused by multi-sentence spans, we concatenate only sentences containing source and events as input to Transformer baselines. These methods outperform the existing reported results and provide strong benchmarks but still perform similarly to a majority class baseline for the TDDMan dataset. Our model shows a significant gain of 8.0 F1 and 8.8 F1 over the BERT baseline on the TDDMan and TDDAuto datasets. Table 2 compares TIMERS to additional rigorous state-of-the-art methods for TimeBank-Dense and MATRES. TIMERS achieves state-of-the-art performance on all four datasets, showing that it successfully handles intra-sentence, inter-sentence, and cross-sentence TLINK pairs through the same architecture.

[1] https://spacy.io/

[2] https://demo.allennlp.org/semantic-role-labeling

[3] https://github.com/microsoft/Recognizers-Text

[4] Implementation used: https://github.com/jiyfeng/DPLP

[5] https://pypi.org/project/discoursegraphs/

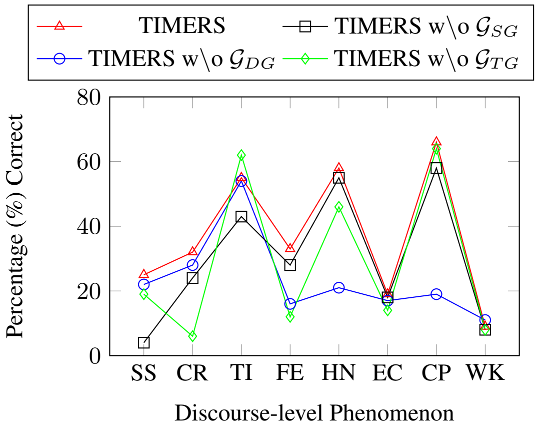


Figure 3: Error analysis on manually annotated discourse-level phenomena in the test set of TDDMan. SS: Single-Sent, CR: Chain Reasoning, TI: Tense Indicator, FE: Future Events, HN: Hypothetical/Negated, EC: Event Coreference, CP: Causal/Prereq, WK: World Knowledge. TIMERS handles CR and CP phenomena but struggles on EC and WK.

3.4 Ablation Study

To assess the contribution of discourse, syntactic, and time-aware graphs, we performed an ablation experiment with different configurations (Table 3). Removing the context encoder significantly degrades performance, indicating that the graph components themselves cannot replace the contextual encoding. Removing any of the graph encoders hurts the model performance, motivating the need for all the constituent graph components. We also analyzed the relative importance of $\mathcal{G}_{DG}$, $\mathcal{G}_{SG}$, and $\mathcal{G}_{TG}$ represented by color shading in the table. The results show that the syntactic graph is least important for document level pairs in TDDMan and TDDAuto, which we believe is due to the longer range dependencies present in this dataset. However, removing the discourse graph for TimeBank-Dense and MATRES datasets leads to the least performance deterioration as inter and intra-sentence pairs do not fully utilize document-level rhetorical relations. TIMERS outperforms the BERT baseline even without $\mathcal{G}_{TG}$, demonstrating its useful in cases where document creation date or timexes cannot be obtained easily.

3.5 Error Analysis

The error analysis results of TIMERS and its ablations for TDDMan are shown in Fig. 3 (the results on TDDAuto are in Appendix Fig.1). The results provide evidence that the syntactic-aware graph ($\mathcal{G}_{SG}$) is most important for relations that can be extracted from a single sentence (SE). The time-aware graph ($\mathcal{G}_{TG}$) plays an important role in improving relationships requiring chain reasoning (multi-hop) and relationship determined by future events. We also note the role of the rhetorical-aware graph ($\mathcal{G}_{DG}$) for modeling future possibility (FE), hypothetical events (HN) and causal conditions for event occurrences (CP). This can be attributed to rhetorical relational features that extract plausible inter-dependencies such as *cause, explanation, contrast* (Lioma et al., 2012). None of the experimented models show improved performance on TLINK pairs which depend on world knowledge (WK) or event coreference (EC).

4 Conclusion

This work presents a neural architecture that utilizes local syntactic features, rhetorical discourse features, and temporal arguments in semantic role labels through a Gated Relational-GCN for document-level temporal relation extraction on TDDiscourse, MATRES, and TimeBank-Dense datasets. Experiments show that TIMERS shows substantial improvement for events that require chain reasoning and causal prerequisite links. Future work will focus on exploring real-world scenarios in which the temporal extraction task suffers from absent or erroneous event and timex annotations. We believe our proposed methods can also be adapted for other languages as well by overcoming possible limitations such as dependency parsing, semantic parsing, Timex normalization for the non-English corpora.

Ethics Statement

This work does not collect or release any new data resource. Moreover, all four of the datasets used in experiments (TDDiscourse, TimeBank-Dense and MATRES) are publicly available and free to use, hence do not intrude user privacy. During the course of this work, no human judgements were exploited nor any user-level data was collected, stored or processed. Our methods do not add to any pre-existing data biases. Potential applications of this work include extracting event timelines from news, contractual documents, and digitizing patient electronic health records. We acknowledge that temporal information extraction finds applications in clinical NLP (Lin et al., 2019; Tourille et al., 2017). Hence, we would like to caution about shortcomings of the proposed system in terms of misclassifications on event pairs requiring real-world common sense reasoning and domain shift.

References

Mohammed Aldawsari, Adrian Perez, Deya Banisakher, and Mark Finlayson. 2020. Distinguishing between foreground and background events in news. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5171–5180, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal. 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.

Fei Cheng and Yusuke Miyao. 2017. Classifying temporal relations by bidirectional LSTM over dependency paths. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Vancouver, Canada. Association for Computational Linguistics.

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.

Tanya Goyal and Greg Durrett. 2019. Embedding time expressions for deep temporal ordering models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4400–4406, Florence, Italy. Association for Computational Linguistics.

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*, abs/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, X. Ren, and Nanyun Peng. 2020a. Deer: A data efficient language model for event temporal reasoning. *ArXiv*, abs/2012.15283.

Rujun Han, Yichao Zhou, and Nanyun Peng. 2020b. 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.

Yangfeng Ji and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13–24, Baltimore, Maryland. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

A. Leeuwenberg and Marie-Francine Moens. 2019. A survey on temporal reasoning for temporal information extraction from text. *ArXiv*, abs/2005.06527.

Artuur Leeuwenberg and Marie-Francine Moens. 2018. Temporal information extraction by predicting relative time-lines. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1237–1246, Brussels, Belgium. Association for Computational Linguistics.

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.

Christina Lioma, Birger Larsen, and Wei Lu. 2012. Rhetorical relations for information retrieval. In *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*, pages 931–940. ACM.

Y. Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Mingyu Derek Ma, J. Sun, M. Yang, Kung-Hsiang Huang, N. Wen, Shikhar Singh, Rujun Han, and Nanyun Peng. 2021. Eventplus: A temporal event understanding pipeline. *ArXiv*, abs/2101.04922.

W. Mann. 1987. Rhetorical structure theory: A theory of text organization.

W. Mann and S. A. Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text Talk*, 8:243 – 281.

Yuanliang Meng, Anna Rumshisky, and Alexey Romanov. 2017. Temporal information extraction for question answering using syntactic dependencies in an LSTM-based architecture. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 887–896, Copenhagen, Denmark. 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.

Arne Neumann. 2015. discoursegraphs: A graph-based merging tool and converter for multilayer annotated corpora. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 309–312, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Qiang Ning, Zhili Feng, and Dan Roth. 2017. A structured learning approach to temporal relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1037, Copenhagen, Denmark. Association for Computational Linguistics.

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, Rujun Han, Nanyun Peng, Matt Gardner, and Dan Roth. 2020. TORQUE: A reading comprehension dataset of temporal ordering questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1158–1172, Online. Association for Computational Linguistics.

Qiang Ning, Hao Wu, and Dan Roth. 2018a. 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.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018b. CogCompTime: A tool for understanding time in natural language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77, Brussels, Belgium. Association for Computational Linguistics.

Yunseok Noh, Yongmin Shin, Junmo Park, A.-Yeong Kim, Su-Jeong Choi, Hyun-Je Song, Seong-Bae Park, and Se-Young Park. 2020. WIRE: an automated report generation system using topical and temporal summarization. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 2169–2172. ACM.

J. Pustejovsky, José M. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, G. Katz, and Dragomir R. Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering*.

Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. 2016. Temporal anchoring of events for the TimeBank corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2195–2204, Berlin, Germany. Association for Computational Linguistics.

M. Schlichtkrull, Thomas Kipf, P. Bloem, R. V. Berg, Ivan Titov, and M. Welling. 2018. Modeling relational data with graph convolutional networks. *ArXiv*, abs/1703.06103.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *ArXiv*, abs/1904.05255.

Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2017. Neural architecture for temporal relation extraction: A Bi-LSTM approach for detecting narrative containers. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–230, Vancouver, Canada. Association for Computational Linguistics.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2906–2919, Florence, Italy. 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.

Zhenyu Zhang, Bowen Yu, Xiaobo Shu, Tingwen Liu, Hengzhu Tang, Wang Yubin, and Li Guo. 2020. Document-level relation extraction with dual-tier heterogeneous graph. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1630–1641, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xinyu Zhao, Shih ting Lin, and Greg Durrett. 2020a. Effective distant supervision for temporal relation extraction. *ArXiv*, abs/2010.12755.

Xinyu Zhao, Shih-ting Lin, and Greg Durrett. 2020b. Effective distant supervision for temporal relation extraction. *arXiv preprint arXiv:2010.12755*.

Yichao Zhou, Yu Yan, Rujun Han, J. Caufield, Kai-Wei Chang, Y. Sun, P. Ping, and W. Wang. 2020. Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. *ArXiv*, abs/2012.08790.

A Experiment Settings

A.1 Node Connections

We detail the node connections present in each graph of our proposed model along with edge attributes in Table 4.

A.2 Edge Relations

Table 6 lists rhetorical relations used in Rhetoric-aware graph G_{DG}in the TIMERS model, along with the definitions as provided by Mann (1987). The weights of the Rhetoric graph G_{DG} are determined based on the RST relations described in this table. Table 7 details the type of relations between timex-timex and DCT-timex nodes of the Time-aware graph G_{TG}.

A.3 Training Setup

Hyperparameter: Hyper-parameters for our model were tuned on the respective validation set to find the best configurations for different datasets. We summarize the range of our model's hyper parameters such as: number of hidden layers in GR-GCN $\{1, 2, 3\}$, size of hidden layers in GR-GCN $\{64, 128, 256, 512\}$, BERT embedding size, dropout $\delta \in \{0.2, 0.3, 0.4, 0.5.0.6\}$, learning rate $\lambda \in \{1e-5, 1e-4, 1e-3, 1e-2, 1e-1\}$, weight decay $\omega \in \{1e-6, 1e-5, 1e-4, 1e-3\}$, batch size $b \in \{16, 32, 64\}$ and epochs (≤ 100).

Contextual Encoder: We used BERT-base-uncased for generating token embedding of size 1x 768. As BERT-base Transformer provides a stronger baseline as compared to RoBERTa, we utilized BERT Transformer for Contextual Encoder in TIMERS architecture. We use the default dropout rate (0.1) on BERT's self attention layers but do not use additional dropout at the top linear layer The output from the Contextual Encoder is a 1-D vector of size 768.

Loss Function and Inference: TIMERS is trained end to end using Binary Cross Entropy loss with Adam optimizer. Across all four datasets, we found the best results correspond with the use of Adam optimiser set with default values $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1e-8$, weight-decay of $5e-4$ and an initial learning rate of 0.001. We evaluate the performance of temporal relation extraction systems in terms of F1, precision and recall score.

Computing Infrastructue: TIMERS is written in PyTorch library and was trained on Nvidia GeForce RTX 2080 GPU. **Average Runtime**: The model takes a maximum of approximately 6,500 seconds to train on either of the four datasets.

Dataset Access Links to download TD-Discourse (Naik et al., 2019) dataset: https://github.com/aakanksha19/TDDiscourse Link to download MATRES (Ning et al., 2018a) dataset: https://github.com/qiangning/MATRES Link to download TimeBank-Dense (Cassidy et al., 2014) dataset: https://github.com/muk343/TimeBank-dense

A.4 Reproducibility

Table 5 lists the range ad best values of the hyperparameters used in TIMERS model for different data settings. We used grid search to choose the best set of training configurations across each dataset. We run 5 rounds of hyper-parameter search trials and report average of observed results.

B Additional Results

We observe from Figure 4 a similar trend to TD-DMan, although with a stronger support for SS, CR, TI and and FE. This is partly due to the fact that TDDAuto was generated automatically (Naik et al., 2019) using weakly annotated time relations. Moreover, 90% of samples in TDDAuto require SS. Hence, TIMERS trained exclusively on TDDAuto performs worse on challenging phenomenon like HN and CP. Consistent with results on TDDMan, TIMERS and its ablations trained on TDDAuto struggle on EC and WK.

Edge	Graph	Source	Target	Directed	Weighted
Document-Sentence Affiliation	Syntactic	Doc Node	Sent Nod	✓	✗
Sentence-Word Affiliation	Syntactic	Sent Nod	Word Node	✓	✗
Sentence-Sentence Adjacency	Syntactic	Sent Nod	Sent Nod	✓	✗
Word-Word Adjacency	Syntactic	Word Node	Word Node	✓	✗
Word-Word Dependency	Syntactic	Word Node	Word Node	✗	✗
DCT-Timex Association	Time	Doc Node	Timex	✓	✓
Timex-Timex Association	Time	Timex	Timex	✓	✓
Predicate-Temporal Argument	Time	Word Node	Timex	✗	✗
RST Discourse	Discourse	EDU	EDU	✓	✓

Table 4: List of node connections in TIMERS.

	Dataset			
Hyperparameters	TDDMan	TDDAuto	MATRES	TB-Dense
Dropout Ratio	0.5	0.5	0.5	0.5
Optimizer	Adam	Adam	Adam	Adam
Input Dimension (Context Encoder)	(n,768)	(n,768)	(n,768)	(n,768)
Input Dimension (Syntactic Graph)	(n,768)	(n,768)	(n,768)	(n,768)
Input Dimension (Time Graph)	(n,256)	(n,256)	(n,64)	(n,64)
Input Dimension (Rhetoric Graph)	(n,768)	(n,768)	(n,768)	(n,768)
Hidden Dimension (GR-GCN)	256	256	64	64
Number of hidden layers (GR-GCN)	1	1	1	1
Hidden Dimension of SpanExt	{256, 64}	{256, 64}	{128, 64}	{128, 64}
Epochs	20	20	20	20
Batch Size	8	8	16	16
Activation Function of Linear layers	ReLU	ReLU	ReLU	ReLU
Dimension of final FCN	[(1792 x r)]	[(1792 x r)]	[(1024 x r)]	[(1024 x r)]
Output Classes	5	5	4	5

Table 5: **Hyperparameters Details:** Training hyperparameters of TIMERS for TDDMan, TDDAuto, MATRES and TB-Dense datasets. n refers to the number of input samples; r refers to the number of total relation classes

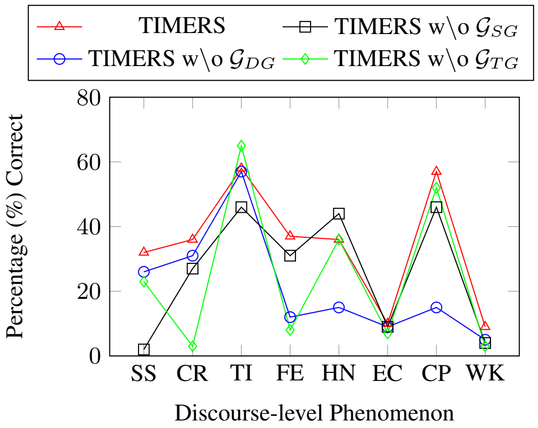


Figure 4: Error analysis on manually annotated discourse-level phenomenon in test set of TDDAuto. SS: SingleSent, CR: Chain Reasoning, TI: Tense Indicator, FE: Future Events, HN: Hypothetical/Negated, EC: Event Coreference, CP: Causal/Prereq, WK: World Knowledge. We observe a stronger support for SS, CR, TI and and FE as compared to TDDMan. TIMERS trained exclusively on TDDAuto performs worse on challenging phenomenon like HN and CP. Consistent with results on TDDMan, TIMERS and its ablations trained on TDDAuto struggle on EC and WK.

Relation Label	Definition
Temporal	Relating to time
Summary	Shorter restatement
Same-unit	Part of the same phrasal unit
Span	Extending to multiple phrasal units
Purpose	Initiation in order to realize a goal
Example	Specific subtypes
Elaboration	Providing additional details
Reason	Justification with intent to defend a stance
Sequence	Subject-matter sequence
Condition	Realization of dependency
Means	Method or instrument to improve likelihood
Consequence	Intended or unintended end goal
Topic	Central idea
Attribution	Contributing factor
Textual Organization	Part of formal text span
Contrast	Opposing phenomenon
Manner	Semantic course of occurrence
Antithesis	Incompatibility due to contrast
Concession	Potential Incompatibility
Explanation	Providing clarification to an established fact
Circumstance	Framework for interpretation

Table 6: RST relations used in Rhetoric-aware graph G_{DG} in TIMERS, with definition as provided by Mann (1987)

Relation Label	Definition
After	TIMEX1 starts after TIMEX2 has ended
Before	TIMEX1 ends before TIMEX2 started
Equal	TIMEX1 is numerically equal to TIMEX2 upto date resolution.
None	One of the timex cannot be extracted or normalized

Table 7: Timex-Timex and DCT-Timex relations used in the Time-aware graph G_{TG}.