Modeling Document-level Temporal Structures for Building Temporal Dependency Graphs

Prafulla Kumar Choubey*
Salesforce Research
pchoubey@salesforce.com

Ruihong Huang
Texas A&M University
huangrh@tamu.edu

Abstract

We propose to leverage news discourse profiling to model document-level temporal structures for building temporal dependency graphs. Our key observation is that the functional roles of sentences used for profiling news discourse signify different time frames relevant to a news story and can, therefore, help to recover the global temporal structure of a document. Our analyses and experiments with the widely used knowledge distillation technique show that discourse profiling effectively identifies distant inter-sentence event and (or) time expression pairs that are temporally related and otherwise difficult to locate[1].

1 Introduction

Grounding all events and time expressions to a reference timeline is fundamental to text understanding. Recently, Yao et al. (2020) proposed a new task and dataset for building temporal dependency graph (TDG)[2]. TDG is based on the notion of narrative time and temporal anaphora, and references each timex to a timex or a meta node and each event to a timex and maybe an event. The reference timex of an event is either the smallest time (when identifiable) that encloses the event or the document creation time (DCT). Similarly, the reference event is selected such that it gives the most precise temporal interpretation for a child event.

Because each event and timex is referenced to only one timex (or additionally an event), identified temporal relations represent the most salient relations that can potentially be used to infer additional temporal relations through transitivity or commonsense reasoning (Yao et al., 2020). This makes identifying reference timex and reference event more challenging, especially when they are mentioned across sentences. Human evaluations by (Yao et al., 2020) also found that identifying the appropriate reference timex and reference event was the most challenging aspect of their annotation.

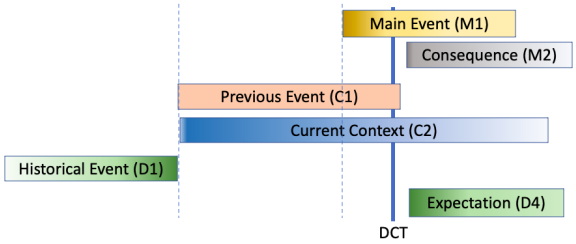

Figure 1: Temporal structures induced by different content types from the News Discourse Profiling.

In this work, we focus on improving cross-sentence reference timex and event mentions identification by exploring discourse-level temporal cues. We choose the news discourse profiling structure (DP) (Choubey et al., 2020). DP classifies sentences in a news document into one of eight content types, defined based on the functional role of a sentence in describing the main news story (Teun A, 1986; Van Dijk, 1988a,b; Choubey et al., 2020), and provides an event-based functional interpretation of sentences. The eight content types include main, consequence, previous event, current context, historical, anecdotal, evaluation and expectation.

As shown in Figure 1, different content types induce different time frames relevant to a news story that can be beneficial for the global interpretation of temporal orders among event and timex mentions. For instance, mentions in historical sentences have a temporal adjacency with other mentions in historical sentences but are likely to be distant from mentions in other content types. Similarly, mentions in previous event sentences may have a temporal adjacency with mentions from one of the previous event, main event or current-context sentences but are likely to be separated from mentions in any

*Work done while at Texas A&M University

[1]Code is available at https://github.com/prafulla77/Discourse_TDG_AACL2022

[2]The dataset was obtained from https://github.com/Jryao/temporal_dependency_graphs_crowdsourcing

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 2: Short Papers), pages 357–365
November 20–23, 2022.

of the historical, expectation or consequence sentences.

We first summarize the distributional association between the position of reference mentions and discourse content types in §2.3. Then, we propose a knowledge distillation-based method to incorporate discourse knowledge into the TDG system. We experiment with the BERT Devlin et al. (2019) and RoBERTa Liu et al. (2019) pre-trained languages models and find that the proposed knowledge distillation-based TDG system is effective in using discourse-level cues and achieves improved performance on identifying cross-sentence reference mentions while retaining performance on the intra-sentence mention pairs.

2 Background and Analysis

2.1 News Discourse Profiling (DP)

Following the news content schemata proposed by Van Dijk (Teun A, 1986; Van Dijk, 1988a,b), DP (Choubey et al., 2020) defines eight content types. Each content type describes the functional role of a sentence in describing the main news event. *Main event* (M1) sentence describes the major events and subjects of the news article. *Consequence* (M2) describes events that are triggered by the main event. *Previous Event* (C1) describes recent events that are a possible cause of the main event. *Current Context* (C2) describes remaining contextual information. *Historical Event* (D1) describes past events that precede the main events in months and years, *Anecdotal Event* (D2) describes unverifiable facts, *Evaluation* (D3) describes opinionated contents from immediate participants, experts or journalists, and *Expectation* (D4) describes speculations or possible consequences of the main or context events.

2.2 Temporal Dependency Graph (TDG)

TDG (Yao et al., 2020) is a directed edge-labeled graph in which each node is either an event, a timex, or a meta node (e.g. document creation time). The reference for each timex/event node is another timex node or a meta node. Optionally, the temporal position of some events can be more precisely determined by referencing them to another event, and thus they can also have a reference event node. For instance, in Figure 2, the event *incident* can only be temporally positioned with respect to the timex *August 23* while the temporal order of event *broke* can be determined with respect to both the timex *later* and the event *occurred*. The edges between event/ timex node pairs are labeled with one of the *overlap*, *after*, *before* and *included* temporal relations while the edges between a timex node and a meta node is assigned a generic *depend-on* label. In this work, we focus exclusively on identifying the reference timex (and event) for each timex (event) without predicting the temporal relations between them.

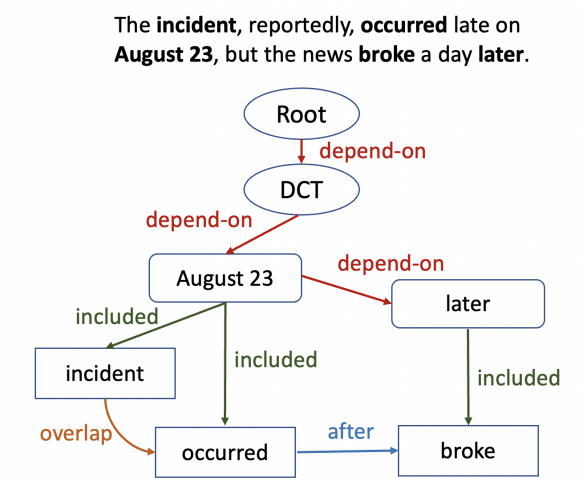

Figure 2: An example TDG.

2.3 Analysis of TDG Structures w.r.t. DP Sentence Types

As illustrated in Figure 1, discourse roles have temporal interpretations that are useful to locate event and timex relations in a document. Therefore, we use the recently proposed discourse profiling system by Choubey and Huang (2021)[3] to assign content type labels to all sentences in the training data and analyze the distribution of reference timex and event mentions across different content types. Note that our analyses are based on a neural network model-predicted discourse content types which are noisy. Additionally, a sentence often contains more than one event and timex mentions and its content type can only provide a broad temporal ordering for constituent mentions.

First, we observe that reference timex for both timex (66% to 100%) and event (54% to 80%) mentions from all content types, except the historical, is majorly the DCT. Further, among the events from non-historical sentences that are not referenced to DCT, we observe that majority (71% to 89%) of them are referenced to a time expression from main,

[3]The discourse profiling system was obtained from https://github.com/prafulla77/Discoure_Profiling_RL_EMNLP21Findings.

current-context, or previous-event sentences that overlaps with the DCT. On the other hand, roughly 66% of the timex mentions in historical sentences are not referenced to any timex mention but to a meta-node. Similarly, over 52% of event mentions in historical sentences are referenced to a timex mention within the same sentence. This is expected given historical sentences describe events from the distant past that are not easily referable to current timex antecedents.

Second, we observe that a significant proportion of cross-sentence event-event relations (45% to 84%) have references in either sentence of the same content type or current context sentences. This can be accounted to the anaphoric nature of TDG representation that only selects reference event which provides the most precise temporal interpretation for a given event. Since sentences with the same content types describe temporally adjacent events, they are conducive to including the most temporally salient related references for all events. The exact distribution of all timex and event mentions across different content types are tabulated in the appendix B.

3 Empirical Evaluations and Results

Based on our observations in §2.3, we perform empirical evaluations to demonstrate the effectiveness of news discourse profiling for building TDG.

3.1 Models

Following recent works on temporal relation identification (Ballesteros et al., 2020) and temporal dependency parsing (Ross et al., 2020), we experiment with pre-trained language models, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019). We model TDG as a ranking problem (Yao et al., 2020), where we add a meta node each for reference timex and event. Then for each event and timex, we obtain the reference timex by selecting the one with the highest score. Similarly, we perform ranking over events to obtain the reference event for each event. To build a TDG from ranking scores, we adopt the technique used by Ross et al. (2020) and iteratively select the highest-ranked reference that does not form a cycle. Within the ranking framework, we develop three models based on each of the BERT and RoBERTa to analyze the role of news discourse structure in building TDG.

Baseline : Given the sentences (x_1^1,..,m_1,..,$x_1^{n_1}$ and x_2^1,..,m_2,..,$x_2^{n_2}$) corresponding to two mentions (m_1 and m_2), we first enclose both mentions in special symbols ($\$m_1\$$ and $@m_2@$) and follow standard language model tokenization step to obtain the context representation sequence (e.g. for RoBERTa, we get <s>,x_1^1,..,$\$m_1\$$,..,$x_1^{n_1}$,</s>, </s>,x_2^1,..,$@m_2@$,..,$x_2^{n_2}$,</s>). Then, we use the pre-trained model to obtain the context representation followed by a linear neural layer to obtain the final score. Note that the context sequence follows the textual order of sentences in a document.

DP-Feature : In addition to the context preprocessing used for the baseline model, it appends special symbols to each sentence corresponding to its discourse content type (e.g. the context for a mention in the main sentence is represented as x_1,..m_1,..x_n,#M1#). Besides that, it mimics the baseline model.

DP-Distillation : It uses the distillation technique (Hinton et al., 2015) to introduce news discourse knowledge into our ranking system. We consider the DP model (Choubey and Huang, 2021) as the teacher network and the *language model* component from the baseline model as the student network. The teacher model generates hard labels for sentences using the *argmax* function. Using the language model, we first obtain embeddings for all sentences in a document and then use a linear neural layer to predict their discourse content types. During training, we perform iterative gradient updates where we first update parameters based on the discourse profiling loss followed by gradient updates based on the temporal ranking loss in each batch. We observe that the order of gradient updates is important. Performing joint gradient updates or switching the order of gradient updates may significantly lower the validation performance.

3.2 Experimental Settings

We use the training, validation and test splits from Yao et al. (2020) for all our experiments. Since our goal here is to evaluate the performance of a model on predicting reference timex and event mentions, we use the gold annotations for event and timex mentions. All three models are trained using AdamW optimizer (Loshchilov and Hutter, 2017) for a maximum of 15 epochs and we use the epoch yielding the best validation performance. We use the batch size of 5 documents and the learning rate of 0.0001 with linear scheduling and warmup steps equivalent to 5 epochs. We search learning rate and warmup steps from [5e-4, 1e-4, 5e-6] and [3,

Model	Valid	Test
Yao et al. (2020)	69.0*	79.0*
BERT		
Baseline	71.90	76.69
DP-Feature	72.04	76.76
DP-Distillation	**72.20**	**78.30**
RoBERTa		
Baseline	74.63	77.26
DP-Feature	74.70	77.30
DP-Distillation	**75.03**	**78.93**

Table 1: Accuracy of different systems on the validation and test datasets. *Results for Yao et al. (2020) are directly taken from the paper and correspond to the single best run.

5, 7] respectively using the baseline model. Then, both the learning rate and warmup steps are kept constant for all models. Each training run takes ∼12 hours for the baseline and DP-feature models and ∼15 hours for the DP-distillation model. RoBERTa or BERT model is fine-tuned during the training. We run each model 3 times with random seeds and report the average performance to reduce the influence of randomness in training.

All experiments are performed on two NVIDIA-RTX-3090-24GB using PyTorch 1.7.1+cu110 (Paszke et al., 2019) and HuggingFace Transformer (v 4.0.1) libraries (Wolf et al., 2019). We use gradient accumulation to fit a batch on 2 GPUs.

3.3 Result and Analysis

Table 1 shows the results from our experiments and the previous best-performing model (Yao et al., 2020). The average accuracy of the baseline model, which relies on the pre-trained RoBERTa (BERT), is 5.63% (2.90%) higher than the best performing neural model from Yao et al. (2020) on the validation dataset. Surprisingly, on the test dataset, our RoBERTa (BERT)-based baseline model achieves 1.74% (2.31%) lower average accuracy.

Next, using discourse content types as a feature in the input sequence brings negligible improvement over the baseline for both RoBERTa and BERT-based models. We suspect that special symbols used to represent each content type are unaware of the temporal associations between different content types. Thus, the DP-feature model is only capable of modeling co-occurrences of different content types with reference event and timex mentions. Additionally, the pre-training of the BERT/ RoBERTa model did not consider special content types symbols which leads to inconsistent interpretation of their corresponding tokens during the pre-training and the fine-tuning steps.

DP-distillation method using the RoBERTa (BERT) model, on the other hand, improves the average accuracy of Baseline by 0.4% (0.3%) and 1.66% (1.61%) on validation and test datasets respectively. Training with the distillation technique enables the transfer of DP knowledge directly from the teacher DP model into the student RoBERTa/ BERT model, unlike the DP-feature model which is unaware of DP knowledge unless specified through features. Further, the DP-distillation model learns to predict content type labels, while being validated over performance on ranking true reference mentions, which provides it with higher flexibility to distill and retain directly relevant knowledge.

Why Discourse Profiling helps? Since DP provides temporal cues at the sentence level, we mainly expect the performance improvement to come from cross-sentence event/ timex pairs. To verify that, we partition our validation and test datasets into three subsets: 1) *intra-sentence* that includes pairs with both given mention and reference mention from the same sentence, 2) *cross-sentence* that includes pairs with given mention and reference mention from different sentences, and 3) *no-parent* that includes mentions which are referenced to a meta node. We compare the RoBERTa based-baseline and DP-distillation models, which perform better than the corresponding BERT-based models, on three data partitions in Table 2. As expected, we found that both the baseline and DP-distillation models achieve comparable performance on the same-sentence subset. For the no-parent subset, we observe higher recall and lower precision for the DP-distillation model. Intuitively, the model learns to link more event and timex mentions to a meta node. Note that timex mentions from historical sentences are majorly linked to a meta node (§2.3), which may be partly responsible for this behavior.

On the cross-sentence subset, we observe consistent improvement on all precision, recall and F1 scores for the DP-distillation model. This is consistent with our hypothesis that discourse profiling can be used to induce document-level temporal structures and help in identifying references for event/timex mentions that require cross-sentence temporal cues.

	Intra-Sentence			Cross-Sentence			No-Parent		
Model	P	R	F1	P	R	F1	P	R	F1
Valid									
Baseline	81.03	**84.66**	**82.8**	70.60	65.36	67.86	**70.03**	79.16	74.3
DP-Distillation	**81.90**	83.66	82.76	**72.00**	**68.86**	**68.76**	67.4	**82.90**	**74.33**
Test									
Baseline	**80.6**	85.86	83.16	75.30	70.56	72.83	**76.20**	80.93	78.43
DP-Distillation	80.53	**86.13**	**83.20**	**79.90**	**71.96**	**75.70**	74.23	**86.16**	**79.70**

Table 2: Precision, recall and F1 scores for RoBERTa-based baseline and DP-distillation models on intra-sentence, cross-sentence and no-parent subsets from the validation datasets.

4 Related Work

Most previous works (Mani et al., 2006; Bethard and Martin, 2007; Kolomiyets et al., 2012; D'Souza and Ng, 2013; Bethard, 2013; Ng et al., 2013; Laokulrat et al., 2013; Mirza and Tonelli, 2014; Choubey and Huang, 2017; Yao et al., 2017; Dai et al., 2017; Yao and Huang, 2018; Ballesteros et al., 2020) treat temporal relation extraction as a pair-wise classification problem and most widely used datasets follow the same pair-wise schema for annotating temporal relations between event/ timex pairs (Graff, 2002; Pustejovsky et al., 2003b,a; Cassidy et al., 2014; UzZaman et al., 2013; Ning et al., 2018). However, as discussed by Zhang and Xue (2018b,a); Ross et al. (2020); Yao et al. (2020), pairwise annotations as well as classification models suffer from quadratic complexity, partial annotations and inconsistent predictions. Recently, Zhang and Xue (2018b) proposed to build a dependency tree (TDT) structure to address the above three problems with pair-wise annotations and modeling and later extended that to temporal dependency graph (Yao et al., 2020). We use the most recent temporal dependency graph dataset that improves the expressiveness of previous TDT datasets (Zhang and Xue, 2018b, 2019) and follow their neural ranking modeling approach. However, different from the previous work, we explore news discourse profiling to explicitly focus on improving the performance of a neural ranking model on cross-sentence event/ timex pairs.

Ng et al. (2013) were the first to show the effectiveness of several discourse analysis frameworks, including rhetorical structure theory (RST) (Mann and Thompson, 1988), PDTB-style discourse relations (Prasad et al., 2008) and topical text segmentation (Hearst, 1994) for temporal relation extraction. Different from the above three discourse structures, discourse profiling is a functional (Webber and Joshi, 2012) structure and has global event-centric interpretations. Secondly, Ng et al. (2013) focused on classifying temporal relations between a given pair of temporally related events. In contrast, our goal is to identify the most salient reference for every event/ timex mention that determines its most precise location on the timeline.

5 Conclusion

We have shown that news discourse profiling can be used to incorporate document-level temporal structures when building temporal dependency graphs. Through analyses, we have shown the distributional association between discourse content types and positions of reference and child mentions. Further, empirical evaluation using the knowledge distillation technique shows that discourse profiling is effective in identifying cross-sentence reference-child mention pairs. In the future, we will explore new linguistics structures and modeling techniques to incorporate document-level temporal structures for building TDG.

6 Acknowledgements

We gratefully acknowledge support from National Science Foundation via the awards IIS-1942918. We would also like to thank the anonymous reviewers for their feedback.

References

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.

Steven Bethard. 2013. ClearTK-TimeML: A minimalist approach to TempEval 2013. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 10–14, Atlanta, Georgia, USA. Association for Computational Linguistics.

Steven Bethard and James H. Martin. 2007. CU-TMP: Temporal relation classification using syntactic and semantic features. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 129–132, Prague, Czech Republic. 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.

Prafulla Kumar Choubey and Ruihong Huang. 2017. A sequential model for classifying temporal relations between intra-sentence events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1796–1802, Copenhagen, Denmark. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2021. Profiling news discourse structure using explicit subtopic structures guided critics. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1594–1605, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Prafulla Kumar Choubey, Aaron Lee, Ruihong Huang, and Lu Wang. 2020. Discourse as a function of event: Profiling discourse structure in news articles around the main event. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5374–5386, Online. Association for Computational Linguistics.

Zeyu Dai, Wenlin Yao, and Ruihong Huang. 2017. Using context events in neural network models for event temporal status identification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 234–239, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

Jennifer D'Souza and Vincent Ng. 2013. Classifying temporal relations with rich linguistic knowledge. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 918–927, Atlanta, Georgia. Association for Computational Linguistics.

David Graff. 2002. The aquaint corpus of english news text.

Marti A. Hearst. 1994. Multi-paragraph segmentation expository text. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2012. Extracting narrative timelines as temporal dependency structures. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 88–97, Jeju Island, Korea. Association for Computational Linguistics.

Natsuda Laokulrat, Makoto Miwa, Yoshimasa Tsuruoka, and Takashi Chikayama. 2013. UTTime: Temporal relation classification using deep syntactic features. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 88–92, Atlanta, Georgia, USA. Association for Computational Linguistics.

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*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 753–760, Sydney, Australia. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Paramita Mirza and Sara Tonelli. 2014. Classifying temporal relations with simple features. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 308–317, Gothenburg, Sweden. Association for Computational Linguistics.

Jun-Ping Ng, Min-Yen Kan, Ziheng Lin, Wei Feng, Bin Chen, Jian Su, and Chew-Lim Tan. 2013. Exploiting discourse analysis for article-wide temporal classification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 12–23, Seattle, Washington, USA. 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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003a. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003b. The timebank corpus. *Proceedings of Corpus Linguistics*.

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.

Van Dijk Teun A. 1986. News schemata. *Studying writing: linguistic approaches*, 1:155–186.

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.

Teun A Van Dijk. 1988a. News analysis. *Case Studies of International and National News in the Press. New Jersey: Lawrence*.

Teun A Van Dijk. 1988b. *News as discourse*. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Bonnie Webber and Aravind Joshi. 2012. Discourse structure and computation: Past, present and future. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 42–54, Jeju Island, Korea. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

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.

Wenlin Yao and Ruihong Huang. 2018. Temporal event knowledge acquisition via identifying narratives. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 537–547, Melbourne, Australia. Association for Computational Linguistics.

Wenlin Yao, Saipravallika Nettyam, and Ruihong Huang. 2017. A weakly supervised approach to train temporal relation classifiers and acquire regular event pairs simultaneously. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 803–812, Varna, Bulgaria. INCOMA Ltd.

Yuchen Zhang and Nianwen Xue. 2018a. Neural ranking models for temporal dependency structure parsing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3339–3349, Brussels, Belgium. Association for Computational Linguistics.

Yuchen Zhang and Nianwen Xue. 2018b. 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, Minnesota. Association for Computational Linguistics.

A Responsible NLP Research Checklist

A.1 Limitations and Risks

Our proposed method relies on news articles' specific functional discourse structure, called news discourse profiling. This limits the applicability of the method to the news domain only. We run all experiments on the dataset in the English language. While we expect the method to work well for other languages, provided we have a dataset/ model for constructing the news discourse profiling structure in the target language, we have not verified this experimentally. Our results are based on the average of 3 runs with random seeds. We do not expect any potential risk from the proposed method.

A.2 Artifacts

We use two publicly available datasets, TDG corpus (Yao et al., 2020) and NewsDiscourse corpus (Choubey et al., 2020), for our experiments and analyses. Our implementations are based on the HuggingFace transformers (Wolf et al., 2019) (Apache license 2.0) and we will release our code under the BSD 3 license.

B Distributional Analysis

	DCT	Meta-node
M1	**86.5**	8.5
M2	**88.9**	4.4
C1	**81.9**	9.0
C2	**79.8**	14.6
D1	30.9	**66.1**
D2	**100.0**	-
D3	**88.8**	8.8
D4	**88.4**	10.9
NA	**66.7**	25.0

Table 3: Distribution of timex and their reference timex mentions, for each content type.

	DCT	Intra-sentence
M1	**58.4**	30.4
M2	**60.1**	17.5
C1	**54.3**	28.7
C2	**63.6**	17.1
D1	34.4	**52.3**
D2	**73.5**	6.0
D3	**80.8**	7.1
D4	**75.6**	15.8
NA	**69.0**	20.0

Table 4: Distribution of event and their reference timex mentions, for each content type.

	M1	M2	C1	C2	D1	D2	D3	D4	NA
M1	39.8	1.1	17.2	29.0	6.5	-	1.1	2.2	3.2
M2	69.7	4.5	7.6	13.6	4.5	-	-	-	-
C1	31.7	-	43.9	10.1	9.4	1.0	3.6	1.0	-
C2	36.8	1.1	17.1	29.2	9.3	-	3.5	1.8	1.3
D1	7.2	-	39.6	10.8	36.0	-	4.5	-	1.8
D2	61.3	-	9.7	19.4	6.5	-	3.2	-	-
D3	27.0	1.9	17.1	31.4	8.8	1.0	5.0	3.6	4.7
D4	39.7	2.9	10.3	22.1	5.9	-	4.4	11.8	2.9
NA	-	-	-	81.8	18.2	-	-	-	-

Table 5: Distribution of event and their reference timex mentions over different content types, when the reference timex is not the DCT. We can see that majority (71% to 89%) of the events from non-historical sentences are referenced to a time expression from main, current-context, or previous-event sentences that overlaps with the DCT.

	M1	M2	C1	C2	D1	D2	D3	D4	NA
M1	43.9	1.8	10.5	21.1	5.3	-	12.3	1.8	3.5
M2	19.6	17.4	6.5	52.2	2.2	-	-	2.2	-
C1	15.8	1.7	35.0	30.8	10.0	0.8	3.3	1.7	0.8
C2	8.7	2.8	7.9	62.1	5.6	0.2	9.2	3.3	0.1
D1	3.4	0.7	4.7	31.8	46.6	-	10.8	2.0	-
D2	57.6	3.0	-	15.2	3.0	-	18.2	-	3.0
D3	1.7	0.6	3.5	30.3	3.5	0.6	54.1	4.5	1.1
D4	5.9	1.7	5.9	33.1	6.8	-	33.1	12.7	0.8
NA	-	0.1	-	0.2	0.1	0.1	0.3	-	0.3

Table 6: Distribution of cross-sentence event and their reference event mentions over different content types, where the event and its reference event are from different sentences. We can see that significant proportion of cross-sentence event-event temporal links (45% to 84%) have references in either sentences of the same content type or current context sentences.