

Effective Distant Supervision for Temporal Relation Extraction

Xinyu Zhao Shih-ting Lin Greg Durrett

Department of Computer Science

The University of Texas at Austin

{xinyuzhao, j07171lin}@utexas.edu gdurrett@cs.utexas.edu

Abstract

A principal barrier to training temporal relation extraction models in new domains is the lack of varied, high quality examples and the challenge of collecting more. We present a method of automatically collecting distantly-supervised examples of temporal relations. We scrape and automatically label event pairs where the temporal relations are made explicit in text, then mask out those explicit cues, forcing a model trained on this data to learn other signals. We demonstrate that a pre-trained Transformer model is able to transfer from the automatically labeled examples to human-annotated benchmarks in both zero-shot and few-shot settings, and that the masking scheme is important in improving generalization.¹

1 Introduction

Temporal relation extraction has largely focused on identifying pairwise relationships between events in text. Past work on annotating temporal relations has struggled to devise annotations schemes which are both comprehensive and easy to judge (Pustejovsky et al., 2003; Cassidy et al., 2014). However, even simplified annotation schemes designed for crowdsourcing (Ning et al., 2018b; Vashishtha et al., 2019) can struggle to acquire high-accuracy judgments about nebulous phenomena, leading to a scarcity of high-quality labeled data. Compared to tasks like syntactic parsing (Bies et al., 2012) or natural language inference (Williams et al., 2018), there are thus fewer resources for temporal relation extraction in other domains.

In this work, we present a method of automatically gathering distantly-labeled temporal relation examples. Unlike traditional distant supervision methods (Mintz et al., 2009), we do not rely on a knowledge base, but instead on heuristic cues that

¹Code and datasets available at: <https://github.com/xyz-zy/distant-temprel>

we will then mask out, forcing the model to make inferences from the remaining context. We explore two types of cues, but focus primarily on events that are anchored to orderable timexes (Goyal and Durrett, 2019). These examples can be collected and labeled using an automatic system (Chambers et al., 2014). By then masking the explicit temporal indicators, a model trained on these examples can no longer learn trivial timex-based rules, but must instead attend to more general temporal context cues. We show that a pre-trained model fine-tuned on this data learns general, implicit cues that transfer more broadly to human-annotated benchmarks. This observation follows a trend of recent work showing pre-trained models’ ability to generalize from synthetic data to natural data (Xu et al., 2020; Marzoev et al., 2020).

We implement our approach with pre-trained Transformer models (Devlin et al., 2019; Liu et al., 2019; Clark et al., 2020) similar to a state-of-the-art temporal relation extraction model from the literature (Han et al., 2019). Our model is able to effectively transfer from a distantly-labeled dataset to the MATRES benchmark (Ning et al., 2018b) when used to supplement a small number of in-domain or out-of-domain samples.

2 Classification Model

Our base classification model consists of a pre-trained Transformer (Vaswani et al., 2017) model with an appended linear classification layer, represented in Figure 1. For the majority of our experiments, we use RoBERTa (Liu et al., 2019), which we found to work better than BERT (Devlin et al., 2019) and ELECTRA (Clark et al., 2020) for domain transfer. We chose a single set of hyperparameters by tuning to match the performance of Han et al. (2019); for details see Appendix A.

A single example consists of an event pair

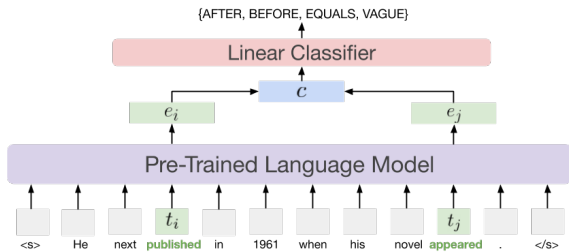


Figure 1: Classification model consisting of a pre-trained transformer model and a linear layer. Event tokens are represented by t_i, t_j and correspond to output embeddings e_i, e_j , which are used by the linear classifier to produce a distribution over labels.

in context, which may be a single sentence or two sentences, and a label from $\{\text{AFTER, BEFORE, EQUALS, VAGUE}\}$, following the annotation scheme in MATRES. Each example is tokenized to yield input tokens $T = [t_1, t_2, \dots, t_n]$, with event tokens $t_i, t_j \in T$. For events consisting of multiple sub-word tokens, we track only the first token position. We use the convention of passing events in text order, where $1 \leq i < j \leq n$.

The language model then produces output embeddings $[e_1, e_2, \dots, e_n]$. For classification, we select the embeddings e_i, e_j corresponding to the event token positions, and combine them into a classification vector, $c = [e_i; e_j; e_i \odot e_j; e_i - e_j]$ where \odot represents elementwise multiplication. Finally, a linear classification layer produces a distribution over the four relation labels. Training is done by maximizing likelihood of labeled samples. We implement this model using PyTorch (Paszke et al., 2019) and pre-trained models from HuggingFace’s Transformers library (Wolf et al., 2019).

We benchmark our classification model by training and evaluating on MATRES. We achieve an F1 of 79.8 with RoBERTa (Liu et al., 2019) and an F1 of 80.3 with ELECTRA (Clark et al., 2020), demonstrating that our model replicates state-of-the-art performance achieved by local models (only considering arcs in isolation), currently 80.3 F1 (Han et al., 2019).

3 Learning from Distant Data

We aim to create a method of automatically gathering high-quality data that can be applied to unlabeled text. To this end, we focus on two techniques identifying explicit temporal indicators. First, we identify single-sentence examples where event pairs are automatically labeled via explicit

discourse connectives. Second, we scrape occurrences of event pairs that are anchorable to timexes which determine their relation. We will see that this second technique is substantially better, and analyze some factors contributing to the performance delta. Although neither technique captures the gamut of phenomena found in human-labeled data, pre-trained models’ generalization capabilities and a masking technique tailored for this setting are two tools that enable effective transfer.

For both techniques, we scrape distant examples from English Gigaword Fifth Edition (Parker et al., 2011). We extract samples from a balance of the different news sources present in the dataset. In both cases, we use the Stanford CoreNLP lexicalized parser (Manning et al., 2014) to generate parse trees for the source text, which can be time-consuming at scale. However, we can pre-filter sentences based on the presence of timexes or target discourse connectives, and so in practice we only rarely need to invoke the parser. Table 9 in the Appendix shows collected data samples, and we describe these two collection methods in more detail below.

3.1 Temporal Connectives

Words like *before*, *after*, *during*, *until*, *prior to*, and others can indicate the temporal status of events in text explicitly. Past work has shown that complex relations can be learned from discourse connectives in non-temporal settings (Nie et al., 2019), so such connectives can be powerful indicators. We focus on *before* and *after* in this work, as these are the most common and straightforward to map to a temporal relation.

To identify connected event pairs, we use the Stanford CoreNLP lexicalized parser (Manning et al., 2014) to produce parse trees. We then search for a related event pair by 1) identifying the connective, 2) finding the closest parent verb phrase, and 3) finding the closest child verb phrase. These become the events for the example. When this identifies modals or auxiliaries, we take the corresponding main verb. The label for the example is simply determined by the *before* or *after* connective. Examples are listed in Appendix E; on inspection, we found this method to be reliable.

3.2 Events Anchored to Time Expressions

Beyond connectives, another cue is the explicit presence of timexes. An example is shown in Figure 2: the years 1951 and 1961 determine the order-

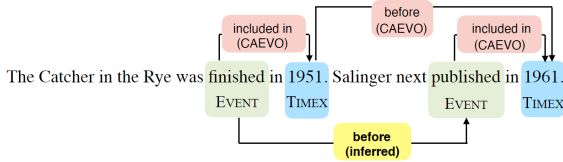


Figure 2: Process of identifying distantly labeled event-event relations using CAEVO. First, events and timexes are identified and two of CAEVO’s sieves are applied. Then, using transitivity, event-event relation is inferred from associated timexes.

ing of their associated events, assuming each event can be appropriately linked to the timex.

We use CAEVO (Chambers et al., 2014) to detect event pairs and link events to timexes, which include explicit datetimes (*January, 1961*), relative times (*tomorrow*) and other natural language indicators (*now, until recently*). This approach, which yields both single- and cross-sentence examples, was explored by Goyal and Durrett (2019), who noisily labeled data to evaluate their timex embedding model.

First, the input document is annotated by CAEVO with events and timexes using its parse trees. Two of its sieves are then applied: the ADJACENTVERBTIMEX sieve identifies events that are anchored to time expressions via a direct path in the syntactic parse tree, then the TIMETIME-SIEVE uses a small set of deterministic rules to label relations between timexes. These two sieves have high precisions, of 0.74 and 0.90, respectively (Chambers et al., 2014). Figure 2 shows the result of applying both sieves. Finally, the system is able to infer the relations between events that are anchored to comparable timexes (i.e. *finished before published*), giving us event pairs usable for training.

The resulting datasets are reasonably balanced between BEFORE and AFTER, with sparse EQUAL examples and *no* VAGUE examples. A more detailed label breakdown is included in Appendix B.

3.3 Example Masking

These distant examples are gleaned from “trivial” indicators in the text, which a model like BERT (Devlin et al., 2019) will overfit to. We observe that our RoBERTa classifier yields 99.8% accuracy on a held out BeforeAfter dataset, and evaluating with the same DistantTimex train/test split of Goyal and Durrett (2019) results in 96.6% test accuracy.

In order to combat this, we *mask* the explicit tem-

poral cues identified by our weak labeling process. Our goal is to induce the model to learn the label from the remaining tokens, including the event instances themselves and the broader context. Masking is performed *prior to* subword tokenization, so each word or timex gets one mask token per word. For our BeforeAfter examples, we simply mask the temporal connective. For our DistantTimex examples, we use the timex tags generated by CAEVO to mask all identified timexes present in the context. This results in masking of not only explicit timexes (e.g. dates, times), but also of natural language timexes (e.g. *previously, recently*). This may occasionally result in “sparse” training examples that have a high proportion of mask tokens. Refer to Tables 7-8 in Appendix E for concrete examples.

In spite of this masking, our model is able to classify distantly-labeled timex examples with 85% accuracy when evaluating on a held-out set, well above a majority baseline. This indicates that there are other temporal cues that the model can use to determine the temporal relation.

4 Results

We evaluate our distant dataset on several axes. (1) In zero-shot, few-shot, or transfer settings, to what extent can this help existing models? (2) How important is each facet of our distant extraction setup? (3) What can we say about the distribution of our data from these techniques? We focus our evaluation on the English MATRES dataset (Ning et al., 2018b), a four-class temporal relation dataset of chiefly newswire data drawn from a number of different sources.²

We found significant variance in model performance in our transfer for small data settings, so most results use average best performance or majority-vote ensembled performance of three randomly seeded models trained in the same setting.

Comparison of Distant Datasets To get the clearest picture of the differences in our distant setups, we first evaluate in a setting of zero-shot adaptation to MATRES. We train on different distantly labeled dataset sizes in order to establish a

²We choose **not** to evaluate on the UDS-T dataset (Vashishtha et al., 2019), treating it solely as a training source. In our experiments, converting from real-valued time span annotations into categorical event-pair labels required dealing with significant disagreement among annotators. Despite trying several resolution strategies, none of our *in-domain* fine-tuned Transformer models performed much better than a majority baseline, indicating high noise or an extremely challenging setting.

Training Data	Split	MATRES		
		P	R	F1
Majority Label	Dev	52.6	60.2	56.1
	Test	50.7	58.6	54.3
MATRES	Dev	77.1	85.5	81.1
	Test	75.1	84.8	79.6
DistantTimex	1k Dev	54.1	61.9	57.7
	1k Test	50.8	58.7	54.5
	5k Dev	60.4	69.1	64.5
	5k Test	61.9	71.5	66.4
	10k Dev	64.0	73.2	68.3
	10k Test	61.5	71.1	66.0
BeforeAfter	10k Dev	53.4	61.1	57.0
	10k Test	51.6	59.7	55.3

Table 1: Results on MATRES after training on distant data, with explicit temporal cues masked. Results are ensembled from three models trained in the same setting.

Labeled Set	Eval	DistantTimex Examples			
		None	1k	5k	10k
MATRES 1k	Avg. Dev F1	64.2	67.4	75.2	75.6
	Avg. Test F1	60.9	66.1	73.6	73.7
	Ens. Dev F1	70.2	74.5	76.7	76.6
	Ens. Test F1	66.5	72.0	75.0	75.5
UDS-T 5k	Avg. Dev F1	68.7	62.8	72.0	70.8
	Avg. Test F1	66.2	60.9	69.5	69.8
	Ens. Dev F1	70.1	64.6	73.2	72.1
	Ens. Test F1	68.2	62.3	70.7	71.8

Table 2: Evaluation results on MATRES when adding automatically collected examples to small amounts of human-annotated training data. Using more DistantTimex data is able to improve performance substantially over not using any (None).

relationship between example quantity and generalization performance. Our results presented in Table 1 show that the DistantTimex data works substantially better than BeforeAfter: with the DistantTimex data, there is a correlation between adding more distantly-labeled examples and increased performance. While both of these rules target particular narrow slices of data, the set with explicit timexes appears to be broader than that with before/after connectives, and hence BERT can learn to generalize better.³

Using Fewer Labeled Examples In a more realistic setting, we assume access to small amounts of pre-existing labeled data, using roughly 10% of existing datasets. Table 2 shows results from

³One possible reason is that explicit indicators like *before* and *after* may be used explicitly to communicate temporal information where it cannot be otherwise inferred, but timexes are often used to communicate more specific details about events where the relation may already be clear.

Training Data	Split	MATRES Test		
		P	R	F1
Distant Timex No Mask	5k	57.9	67.0	62.1
Distant Timex Masked	5k	61.9	71.5	66.4
BeforeAfter No Mask	5k	51.5	59.5	55.2
BeforeAfter Masked	5k	51.6	59.7	55.3

Table 3: Transfer comparison for data with and without masking. Results are ensembled from three models trained in the same setting.

MATRES	DistantTimex	BeforeAfter	UDST 5k
said 16.0%	won 2.02%	was 2.60%	is 5.8%
killed 1.2%	died 1.69%	said 2.08%	was 3.3%
found 1.0%	said 1.67%	came 1.70%	have 2.9%
says 0.9%	began 1.41%	is 1.06%	are 2.5%
told 0.8%	joined 1.10%	began 0.99%	be 2.2%
called 0.8%	took 1.10%	made 0.85%	get 1.5%
reported 0.7%	set 1.04%	have 0.78%	had 1.4%
saying 0.7%	killed 1.04%	left 0.77%	know 1.4%
say 0.7%	born 0.96%	had 0.77%	do 1.1%
was 0.6%	held 0.93%	be 0.73%	go 1.9%

Table 4: Top 10 events for each dataset as a percentage of total event mentions.

evaluating on MATRES using small amounts of either MATRES or UDS-T data in conjunction with our distant data; three models are randomly initialized and trained for each setting. In both settings, adding distant data improves substantially over just using the in-domain MATRES data, and the best model performance is only around 4 F1 worse than the in-domain MATRES results using the entire train set. We also show that this data can stack with data from UDS-T (Vashishtha et al., 2019) and improve transfer over raw UDS-T. This is in spite of very different event distributions between UDST and MATRES and a complete lack of examples of VAGUE relations during training.

Effect of Masking In Table 3, we test the effect of masking on model generalization. We train our model on the collected distant examples with and without masking, and report ensembled evaluation results on the MATRES test set. Our comparison shows that masking causes an increase in generalization for DistantTimex examples, but little change in BeforeAfter transfer, which still performs similar to the majority baseline.

Understanding the data distribution We report the most frequent events from each dataset in Table 4. MATRES is highly focused on reporting verbs, but the distant data has a much flatter distribution. BeforeAfter features more light verbs

whereas DistantTimex features events with more complex semantics; possibly the model can learn more regular and meaningful patterns from such data, or relevant cues from a more similar event distribution (than found in BeforeAfter). We present event-label tuples in Appendix D.

5 Related Work

There is little direct prior work on using this kind of distant supervision for temporal relation extraction. Past work has studied automatic extraction of typical inter-event orderings (Chklovski and Pantel, 2004; Ning et al., 2018a; Yao and Huang, 2018) to aid downstream temporal tasks, but these approaches represent events as single words (predicates) taken out of context, so the knowledge they can capture is limited. The commonsense acquisition method of Zhou et al. (2020) learns more sophisticated information, but more about unary properties of events (typical time, duration) rather than relational knowledge. Lin et al. (2020) achieve a somewhat similar goal, but make a strong assumption about narrative-structured corpora and do not evaluate on in-context temporal relation extraction.

Our technique does not use a knowledge base like classic distant supervision methods (Mintz et al., 2009). However, because we eventually mask out the explicit temporal indicators, we are still using temporal information “external” to the final example to derive the label, hence why we invoke this term. A related concept is the idea of labeling functions (Ratner et al., 2016; Hancock et al., 2018), which are used to automatically construct training data for new domains. However, to our knowledge, these techniques have not been applied to temporal relation extraction, nor used in conjunction with masking as we do.

6 Discussion

We use explicit temporal cues to automatically identify examples of temporal relations between events in text. By masking these trivial features, a pre-trained Transformer model can learn from the remaining context and generalize to human-annotated benchmarks. Comparing performance for two distant labeling methods—using discourse connectives and linking events to time expressions—indicates that richer temporal cues exist in the second case. The scope of identified time expressions encompasses both explicit datetimes and natural language indicators (“now”, “recently”, etc.).

While datetimes may be more common in news and historical articles, relative time expressions are present in diverse domains such as literature and colloquial texts. Where such indicators exist, our approach may be used to automatically collect distantly labeled temporal relations. More broadly, we believe that this label-and-mask paradigm could be used to collect targeted training data for a variety of NLP tasks.

Acknowledgments

Thanks to the anonymous reviewers for their helpful comments. This material is also based on research that is in part supported by the Air Force Research Laboratory (AFRL), DARPA, for the KAIROS program under agreement number FA8750-19-2-1003. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Air Force Research Laboratory (AFRL), DARPA, or the U.S. Government.

References

- Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. *English Web Treebank*. LDC2012T13. Linguistic Data Consortium, Philadelphia, PA.
- 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.
- Timothy Chklovski and Patrick Pantel. 2004. *VerbOcean: Mining the web for fine-grained semantic verb relations*. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 33–40, Barcelona, Spain. Association for Computational Linguistics.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. *Electra: Pre-training text encoders as discriminators rather than generators*. In *International Conference on Learning Representations*.

- 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, Qiang Ning, and Nanyun Peng. 2019. [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*, pages 434–444.
- Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. [Training classifiers with natural language explanations](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, Melbourne, Australia. Association for Computational Linguistics.
- Shih-Ting Lin, Nathanael Chambers, and Greg Durrett. 2020. Conditional Generation of Temporally-ordered Event Sequences. *arXiv cs.CL 2012.15786*.
- 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](#). *CoRR*, abs/1907.11692.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. [The Stanford CoreNLP Natural Language Processing Toolkit](#). In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.
- Alana Marzoev, Samuel Madden, M. Frans Kaashoek, Michael Cafarella, and Jacob Andreas. 2020. [Unnatural language processing: Bridging the gap between synthetic and natural language data](#).
- Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. [Distant supervision for relation extraction without labeled data](#). In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.
- Allen Nie, Erin Bennett, and Noah Goodman. 2019. [DisSent: Learning Sentence Representations from Explicit Discourse Relations](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.
- Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018a. [Improving Temporal Relation Extraction with a Globally Acquired Statistical Resource](#). 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 841–851, New Orleans, Louisiana. Association for Computational Linguistics.
- Qiang Ning, Hao Wu, and Dan Roth. 2018b. [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.
- Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. *English Gigaword Fifth Edition*. LDC2011T07. Linguistic Data Consortium, Philadelphia, PA.
- 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.
- James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The TimeBank corpus. *Proceedings of Corpus Linguistics*.
- Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. 2016. Data programming: Creating large training sets, quickly. NIPS’16, page 3574–3582, Red Hook, NY, USA. Curran Associates Inc.
- 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.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinukasz Kaiser, and Illia Polosukhin. 2017. Attention

is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*,.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Silei Xu, Sina Semnani, Giovanni Campagna, and Monica Lam. 2020. AutoQA: From databases to QA semantic parsers with only synthetic training data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 422–434, 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.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589, Online. Association for Computational Linguistics.

A Model Implementation

As mentioned previously, we selected a set of hyperparameters by fine-tuning to approximate the performance of recent state of the art on MATRES, achieving an F1 of 79.8 with RoBERTa (base) and 80.3 with ELECTRA. Specifically, we arrived at a learning rate of 2e-5 with a warmup proportion of 0.1. Our batch size varied from 16-25 based on hardware.

Full implementation can be found at <https://github.com/xyz-zy/distant-temprel>

B Label Composition Across Datasets

Table 5 compares the label distribution of our distant data against the human-annotated MATRES dataset. In comparison, our datasets are lacking

	BEFORE	AFTER	EQUAL	VAGUE
MATRES	34.5%	50.2%	3.5%	11.8%
DistantTimex	67.4%	31.4%	1.3%	–
BeforeAfter	50.7%	49.3%	–	–

Table 5: Training data label distribution. Both of our distant datasets contain 10k examples; the MATRES training set contains 9.7k examples.

Training Data	Split	MATRES			–Vague
		P	R	F1	Acc.
Majority Label	Dev	52.6	60.2	56.1	–
	Test	50.7	58.6	54.3	–
MATRES	Dev	77.1	85.5	81.1	85.5
	Test	75.1	84.8	79.6	84.8
DistantTimex	1k Dev	54.1	61.9	57.7	61.9
	1k Test	50.8	58.7	54.5	58.7
	5k Dev	60.4	69.1	64.5	69.1
	5k Test	61.9	71.5	66.4	69.2
	10k Dev	64.0	73.2	68.3	73.2
	10k Test	61.5	71.1	66.0	71.1
BeforeAfter	10k Dev	53.4	61.1	57.0	61.1
	10k Test	51.6	59.7	55.3	59.7

Table 6: Expanded view of results on MATRES, comparing performance on only on {BEFORE, AFTER, EQUALS} examples (“– Vague”) versus the entire eval set. Presented results are majority-vote ensembled from three models trained in the same setting.

in VAGUE relations (which have the lowest IAA in MATRES, and are particularly difficult to resolve) but emphasize the two most prominent classes. In few-shot settings, our model sees and trains on VAGUE relations from MATRES.

C Performance without VAGUE Examples

The VAGUE label is defined to express indeterminacy and also has the lowest inter-annotator agreement in the MATRES dataset. These examples are also relatively more difficult for models to learn. Tables 6 and 7 present an expanded view of our results, adding the evaluation accuracy on only {BEFORE, AFTER, EQUALS} examples. As expected, we observe that model performance increases across the board under this evaluation.

D Most Common Event-Label Tuples

Table 8 presents a comparison of the most common (event1, event2, label) tuples across datasets. In MATRES, the most common tuples are largely (event, “said”, BEFORE) events. The DistantTimex data features many examples of same-verb pairs

Labeled Set	+Distant	Full Test Set		– Vague
		Acc.	F1	Acc.
MATRES 1k	1k	67.0	72.0	77.5
	5k	69.2	75.0	79.1
	10k	70.1	75.5	80.8
UDS-T 5k	1k	58.1	62.3	67.1
	5k	65.9	70.7	76.2
	10k	67.9	71.8	78.5

Table 7: Comparison of majority-vote ensembled performance on {BEFORE, AFTER, EQUALS} examples (“– Vague”) versus performance on the entire test set. Performance is higher without vague examples, and increases with the number of DistantTimex examples added.

MATRES	
(‘said’, ‘said’, BEFORE)	1.69%
(‘said’, ‘said’, VAGUE)	0.67%
(‘said’, ‘said’, AFTER)	0.25%
(‘said’, ‘told’, BEFORE)	0.21%
(‘called’, ‘said’, BEFORE)	0.21%
(‘killed’, ‘said’, BEFORE)	0.16%
(‘found’, ‘said’, BEFORE)	0.15%
(‘added’, ‘said’, BEFORE)	0.15%
(‘found’, ‘said’, AFTER)	0.14%
(‘declined’, ‘said’, BEFORE)	0.14%
DistantTimex	
(‘won’, ‘won’, BEFORE)	0.44%
(‘died’, ‘died’, BEFORE)	0.26%
(‘annexed’, ‘seized’, BEFORE)	0.26%
(‘born’, ‘graduated’, BEFORE)	0.25%
(‘crashes’, ‘killed’, AFTER)	0.20%
(‘killed’, ‘torched’, AFTER)	0.19%
(‘crashed’, ‘crashed’, AFTER)	0.17%
(‘died’, ‘married’, BEFORE)	0.16%
(‘end’, ‘remove’, BEFORE)	0.14%
(‘died’, ‘killed’, BEFORE)	0.14%

Table 8: Top 10 event-event-relation tuples per dataset as a percentage of total event mentions.

(*won-won, died-died*). These examples frequently come from sentences or sentence pairs discussing related events of the same type, using dates to contrast them.

E Distant Examples

Table 9 presents a sample of our distantly labeled data. Examples (a-b) show that our BeforeAfter parsing scheme can correctly identify linked events across sentence spans. Examples (c-d) display a variety in parsed syntactic structures that link events to timexes.

Distant Example Masking Tables 10 and 11 present examples of our masking scheme on BeforeAfter and DistantTimex examples respectively.

Notably in Table 11, multi-word timexes result in one mask token per word. All identified timexes in the examples are masked, even if they are not directly linked to the events in consideration.

	Example Text	Heuristic	Label
(a)	A child prodigy, he studied music at the Julliard School of Music <u>before</u> turning to art.	BeforeAfter	BEFORE
(b)	Hill’s support came only a few days <u>after</u> former Prime Minister Paul Keating made a speech at the University of New South Wales in which he revived his campaign for a republic.	BeforeAfter	AFTER
(c)	<u>1903</u> : General Electric introduces the first light set for the public sale for \$12, then the average weekly wage of a typical American worker. <u>1962</u> : General Electric designs the National Christmas Tree for the first time.	DistantTimex	BEFORE
(d)	Batista’s EBX, the holding company for his five Rio-based companies, increased its workforce fivefold since <u>2006</u> to 2,000 employees. State-run oil producer Petroleo Brasileiro... plans to add 6,000 more by <u>2013</u> .	DistantTimex	BEFORE
(e)	It took until <u>recently</u> , the official said , for everyone to realize “it didn’t work.”	DistantTimex	BEFORE

Table 9: Weakly labeled data examples, with the identifying heuristics underlined. Examples (a-b) show different instances of labeling based on “before/after” mentions. Examples (c-e) show instances of labeling event pairs based on anchoring time expressions, which range from datetimes to natural language indicators (e.g. “recently”). While other timexes are identified in the examples shown, only the anchoring expressions for the induced labels are indicated.

	Example Text	Masked	Label
(a)	A child prodigy, he studied music at the Julliard School of Music <u>before</u> turning to art.	A child prodigy, he studied music at the Julliard School of Music [mask] turning to art.	BEFORE
(b)	Hill’s support came only a few days <u>after</u> former Prime Minister Paul Keating made a speech at the University of New South Wales in which he revived his campaign for a republic.	Hill’s support came only a few days [mask] former Prime Minister Paul Keating made a speech at the University of New South Wales in which he revived his campaign for a republic.	AFTER

Table 10: Masking of of weakly labeled data examples identified by before/after mentions.

	Example Text	Masked	Label
(a)	Batista’s EBX, the holding company for his five Rio-based companies, increased its workforce fivefold since <u>2006</u> to 2,000 employees. State-run oil producer Petroleo Brasileiro has hired 22,000 employees in the past six years, ... and plans to add 6,000 more by <u>2013</u> .	Batista’s EBX, the holding company for his five Rio-based companies, increased its workforce fivefold since [mask] to 2,000 employees. State-run oil producer Petroleo Brasileiro has hired 22,000 employees in [mask] [mask] [mask] [mask], ... and plans to add 6,000 more by [mask].	BEFORE
(b)	It took until recently, the official said , for everyone to realize “it didn’t work.”	It took until [mask], the official said , for everyone to realize “it didn’t work.”	BEFORE

Table 11: Masking of of weakly labeled data examples identified by anchoring timexes.