Generic Temporal Reasoning with Differential Analysis and Explanation

Yu Feng[1*] **Ben Zhou**[2] **Haoyu Wang**[2] **Helen Jin**[2] **Dan Roth**[2]
[1]Duke University [2]University of Pennsylvania
yu.feng344@duke.edu {xyzhou, why16gzl, helenjin, danroth}@seas.upenn.edu

Abstract

Temporal reasoning is the task of predicting temporal relations of event pairs with corresponding contexts. While some temporal reasoning models perform reasonably well on in-domain benchmarks, we do not understand said[Annie: ** wrong word? **] systems' generalizability as they rely heavily on end-to-end mechanisms. In this work, we introduce a novel framework and dataset named TODAY that aims to bridge this gap and evaluate if a temporal model can make correct predictions based on the right reasons. TODAY relies on differential analysis, which adds an additional sentence to the context and checks if systems can identify how this subtle change may affect the temporal relation of a given event pair. For each context change, we also annotate human explanations of how the change is connected with the temporal prediction. We view TODAY's formulation as a close proxy to evaluating system explanations, which is the most intuitive way to sanity-check models' reasons yet impossible to evaluate automatically. With the help of TODAY, we show that existing models, including GPT-3, drop to almost random guessing on our benchmark, suggesting that they heavily rely on spurious information rather than proper reasoning for temporal predictions. We show that TODAY's supervision style, along with its explanation annotations, can be used to solicit robust distant supervision and encourage models to learn to use the right signals during training. As a result, TODAY contributes to more generic and reliable temporal reasoning systems that outperform baseline systems across several benchmarks.

1 Introduction

Temporal relation extraction is traditionally viewed as an information extraction task, where a model uses explicit temporal signals such as "before" to identify the temporal order of events. While these models have contributed to many downstream pipelines, they are not enough for more complicated tasks such as timeline generation, where most event pairs do not come with explicit clues. These implicit temporal relation extractions require temporal reasoning that relies on common sense and semantic understanding of the context. In recent work, a popular approach to address these predictions is to finetune pre-trained language models with annotated supervision data. Unfortunately, existing benchmarks' supervision and evaluation data often come from the same distribution, so models with millions of parameters can easily exploit spurious signals and annotation artifacts. This means that the community has almost no insight into the generalizability of current temporal reasoning models.

In this work, we bridge this gap with a novel framework that evaluates whether a temporal reasoning model is getting the predictions right with the right reasons. Our intuition is to ask models to *explain* temporal relation predictions since the most viable way for humans to demonstrate a real understanding of these problems is by providing satisfactory explanations to another human being. While the motivation is sound, automatically evaluating the plausibility of model explanations is extremely difficult. As a result, our framework employs an approximation of explanations, which we call temporal differential analysis [Haoyu: bold?]. Under this setting, we select event pairs where the temporal relations are not 100% deterministic based on the context, meaning that both before/after relations can be possible if additional information regards [Haoyu: in regard to] the context is given. Then, we annotate two hypothetical changes to the context for each such event pair, one change makes the event pair's temporal relation "more before", and the other change makes it "more after". Each hypothetical change is also annotated with human explanations of why the change affects

[*]Work partly done when visiting UPenn.

the temporal relation. We collect [BZ: NUMBER HERE] such event pairs with a rigorous human annotation pipeline and call the resulting dataset TODAY (**tempo**ral **d**ifferenti**a**l anal**y**sis). As the full name suggests, TODAY performs a differential analysis on temporal reasoning. If a model is generic enough to provide proper explanations for its temporal decisions, it can also distinguish minor [Haoyu: subtle] context changes and understand how each change will affect the distribution of temporal relations.

[BZ: Provide a figure with an example here.] [Annie: ** this example should also demonstrate what is spurious information **]

We find that models that achieve relatively high in-domain test performances are brittle and demonstrate minimal capabilities for differentiating subtle context changes that affect temporal relations. For example, T5-large achieves 66.2% binary accuracy on TRACIE but only gets 52.3% on TODAY, which is barely above random guessing. Large language models (LLM) such as GPT-3 also fails on this task, getting only [BZ: number here] binary accuracy in a few-shot setting. To mitigate this gap, we propose a general technique that uses temporal explanations that TODAY annotates. Specifically, we argue that explanations of temporal relations are a great proxy for both understanding the evaluating the understanding of temporal reasoning[Haoyu: ?]. We show models trained with TODAY's task formulation and explanation annotation are better at perceiving cross-dataset supervision, and achieve superior performances on multiple datasets with a single model. This is one step closer towards generic temporal reasoning.

We also find that while LLMs are not good enough for temporal differential analysis, they sometimes produce reasonable explanations towards a given temporal relation. Based on this finding, we design a pipeline that automatically collects supervision signals. The pipeline starts with giving GPT-3 an instance from TODAY and a hypothetical temporal relation, then we use GPT-3 to generate several explanations, and finally we train an explanation verifier based on our human annotation and use that to select auto explanations that are more likely to be plausible. We show that adding such data from GPT-3 further boosts the performance across our benchmarks.

Our contribution in this work is threefold. 1) We design a novel evaluation framework and collect a new dataset TODAY that uses differential analysis to test whether systems can perform temporal reasoning with the right reasons; 2) We show that the TODAY supervision, especially the use of explanations contribute towards a generic temporal reasoning model; 3) We use LLMs to generate pseudo explanations and filter them with a novel explanation verification model, and show that such distant supervision signals are helpful.

[BZ: Todo for Ben: mention that explanation can bridge the gap, but expensive, so we use GPT.]

2 Related Work

Temporal Reasoning Models. [Haoyu: Add a sentence here] To improve the quality of event representations, Mathur et al. (2021) embrace rhetorical discourse features and temporal arguments; Trong et al. (2022) select optimal context sentences via reinforcement learning to achieve SOTA performances; while Liu et al. (2021); Mathur et al. (2021); Zhang et al. (2022) employ graph neural networks to avoid complex feature engineering. From the learning perspective, Ning et al. (2018a), Ballesteros et al. (2020), and Wang et al. (2020) enrich the models with auxiliary training tasks to provide complementary supervision signals, while Ning et al. (2018b), Zhao et al. (2021) and Zhou et al. (2021) bring into play distant supervision from heuristic cues and patterns.

3 Dataset

[Annie: ** we may need to uniform name: event pair, context **]

[Annie: ** We haven't defined clearly what is the superious correlation in the temporal reasoning models **] In this section, we introduce the evaluation framework and the collection process of TODAY.

3.1 Task overview

[BZ: Below is my version for Sec3.1. original ver. commented.] The TODAY dataset and its overall framework is designed to evaluate systems' ability to make temporal predictions with plausible reasons. Existing datasets, including MATRES ()

The additional sentence together with the explanation sentence can serve as the approximate explanation for an instance. This formulation in general instantizes the temporal explanation structure framework for common sense category but still provides similar information. It also prevents

Example
Paragraph 1: Dave was a scientist. Dave wanted to make a great scientific discovery. Dave worked with algae to make electricity. Dave discovered he could make electricity with algae! Dave was awarded for his great discovery.
Paragraph 2: Dave wanted to make a great scientific discovery. Dave worked with algae to make electricity. Dave discovered he could make electricity with algae! Dave was awarded for his great discovery.
Additional sentence: Dave was a scientist.
Statement: Dave applied for a grant for his project starts before Dave worked with algae to make electricity
Why does the additional sentence make the relation more before?: The additional sentence implies Dave was a scientist and normally a scientist has to apply for a grant before he starts the project.

Table 1: An example for differential Analysis

models from using too many spurious information since models are forced to explain the subtle context change instead of the entire passage where they could cheat more. Specifically, this formulation can assure an **faithful** and **plausible** explanation (). A precise additional sentence can be regarded as a faithful explanation as it implies the prediction label. An accurate explanation for the additional sentence corresponding to the label can be regarded as a plausible explanation as it verifies how well an explanation (the additional sentence) supports a predicted label.

We conduct a pilot human study for this formulation and find out it is easy to annotate and achieve substantial improvement over the explanation quality compared with the above explanation structured framework. We therefore adopt this formulation as approximate temporal explanation and instances following this formulation are created through a multi-stage annotation process as detailed below.

3.2 Dataset Construction

The TODAY dataset is constructed in three steps: 1) Implicit event generation 2) Crowdsourcing explanation annotation 3) Crowdsourcing evaluation. Step 1 is implemented with GPT-3 prompting and all the remaining steps are implemented with the Amazon Mturk platform with qualification exams.

Each TODAY instance contains 1) a context (or premise) consisting of a sequence of explicit narrative events; 2) a statement that describe a temporal relation of either starts before,starts after for an event pair that happened and are inferable and relevant to the story (One implicit event and one explicit event); 3) an additional first sentence for the context to make the statement with a certain temporal relation more likely to hold true; 4) an explanation sentence for the additional first sentence to explain why adding the sentence makes the statement more likely to hold true.

Implicit Event Generation. We randomly sample short stories from the ROCStories dataset (). For each story, we require GPT-3 to generate an implicit event phrases that are not explicitly mentioned by the given context, but are inferable and relevant. We prompt the model with several (context, event) triplets followed by a context instance which we expect the model to generate an implicit event.

Crowdsourcing explanation annotation. Given an original context and a temporal relation of an event pair, instead of giving a worker another different context, we ask the worker to first provide an additional sentence as a first sentence that makes the statement that contains the temporal relation more likely to hold true. The worker is then asked to explain why adding the sentence makes the relation more likely to hold true. This formulation can be easier to implement. Specifically, we annotate two hypothetical changes to the context for each event pair by assigning the temporal relation of starts before and starts after for the same context and event pair respectively. As a result, for a given context and an event pair, we annotate two opposite instance, one change makes the event pair's temporal relation "more before", and the other change makes it "more after".

Crowdsourcing evaluation. Given that existing automatic metrics often do not correlate well with human judgements of explanation quality (), we conduct crowdsourcing evaluation. We present workers with an original context, a statement including the temporal relation of an event pair, an additional first sentence and its corresponding explanation sentence. We then ask them to decide if the additional sentence makes the statement of the temporal relation more likely to hold true and if the explanation sentence explains why the additional sentence makes the statement more likely to be true, collecting 2 annotations per data point. The instances that are accepted by 2/2 workers will be assigned a third worker.

3.3 Splits and Analysis

We gather 1k instances that are accepted by 3/3 workers to form the testing set. The 1241 instances that are accepted by at least 1 workers are put into the training set.

4 Modeling

[Annie: ** Should we mention what is the input data like? **] We use a pre-trained sequence-to-sequence model as our base model and finetune the model on the temporal reasoning benchmarks TRACIE (Zhou et al., 2021) or/and MATRES () together with TODAY's task formulation and explanation annotation. Each training instance of the model includes an instance from TRACIE(or MATRES) and an instance from TODAY. We compute a two-class cross-entropy loss ℓ_{CE} with logits for the instance from TRACIE. For the instance from TODAY, we adopt the margin ranking loss function to explicitly learn from the explanation that the probability change of the correct temporal relation direction should be larger than the incorrect direction with the guidance of the additional sentence and explanation towards a particular temporal relation[Annie: ** seems awkward narratives here, need proofread **],

$$\ell_{MR} = \max(0, \epsilon + p_g - p_{o_g}) \\ +\max(0, \epsilon + p_{o_w} - p_w), \quad (1)$$

where p_g / p_w is the logit of the instance with additional sentence and explanation sentence given the corresponding correct / incorrect temporal relation and p_{o_g} / p_{o_w} is the logit of the instance without additional sentence and explanation sentence given the correct / incorrect temporal relation. ϵ is a margin separating the logits. The final loss function is ,

$$\ell = \ell_{CE} + \alpha \ell_{MR} \quad (2)$$

where α reduces the two losses into the same scale.

4.1 LLM Distant Supervision

[Annie: ** we could mention our explanation formualtion is especially suitable for LLM, as it also adopt the chain-of-thought idea **] Collecting high-quality written explanations to serve as supervision is difficult and expensive. Recent progress in prompting large language models (LLMs) provides a potentially promising alternate **??**. These days, LLMs have proven to be generalist agents that work surprisingly effective across a range of NLP tasks with the in-context learning paradigm. While LLMs are not good enough for specific tasks like temporal differential analysis, with a proper prompt, they sometimes produce reasonable explanations towards a given temporal relation. We therefore distill GPT-3 by creating a training pipeline that combines GPT-3 with weak explanation verifiers to solicit a large set of automatic reasonable explanations from GPT-3.

4.2 Few-shot prompting for explanations

We adopt the same way in the dataset construction to generate implicit events given stories from ROCStories. We then prompt GPT-3 with several (context, statement, additional sentence, explanation sentence) triplets followed by an unexplained context, statement instance for which we expect the model to generate an additional sentence and explanation sentence. We demonstrate an example in table 2.

Rule-based filter. GPT-3 may cheat by generating an almost identical sentence from the context as the additional sentence or generating the exact statement that explicitly mentions the temporal relation as the explanation sentence. We apply Sentence-BERT (Reimers and Gurevych, 2019) to perform a sentence similarity test to directly filter the GPT-3 generated instances that fall in this two categories.

4.3 General explanation verifier

Concretely, given the instance, the gold label temporal relation, and a generated explanation, the explanation verifier predicts whether the explanation is acceptable. We adopt the abovemenioned model finetuned with TRACIE(or/and MATRES) and TODAY as the general explanation verifier. The instance with correct generated explanation, i.e., the correct additional sentence and correct explanation sentence towards the gold temporal relation will be accepted by the verifier.

4.4 Additional verifiers

Since the general verifier is trained with an additional sentence together with a corresponding explanation sentence, it will be hard for the general explanation verifier to decide whether the instance is acceptable if the instance is partially correct. We therefore propose two additional verifiers, an additional sentence verifier to filter a wrong additional sentence and and an explanation sentence verifier to filter a inappropriate explanation sentence.

Let's add a sentence as the first sentence of the paragraph to let the statement more likely to hold true and explain why.
Paragraph: Tara always wanted jewelry. Her birthday was coming up. Test went to the store. He gave her a really nice necklace She adored him for the gift.
Statement: Test was being a good friend **starts before** he give her a really nice necklace
Add what sentence as the first sentence of the paragraph and why is the statement more likely to hold true?
Test and Tara always hanged out together.
This makes the statement true because normally people will only hang out frequently with their friends and friends will send each other gifts on their birthdays.
###
Paragraph: Tara always wanted jewelry. Her birthday was coming up. Test went to the store. He gave her a really nice necklace She adored him for the gift.
Statement: Test was being a good friend **starts after** he give her a really nice necklace
Add what sentence as the first sentence of the paragraph and why is the statement more likely to hold true?
Test had always had the biggest crush on his classmate Tara even though she didn't talk to him much.
This makes the statement true because it indicates that Test and Tara's relationship wasn't close prior to Test giving Tara the gift.
###
Paragraph: Tim's tooth was hurting like crazy. He could barely eat or drink. His dentist took a look around in his mouth. One of his teeth was rotten. Once the tooth was pulled, Tim felt fine.
Statement: Tim scheduled an appointment with his dentist starts before his tooth was hurting like crazy
Add what sentence as the first sentence of the paragraph and why is the statement more likely to hold true?

Table 2: A sample prompt with an instance for two hypothetical changes to make the event pair's temporal relation "more before" or "more after".

Additional sentence verifier. We adopt the same training model paradigm as the general explanation verifier. Instead of consuming a complete TODAY instance together with a TRACIE(or MATRES) instance as a input instance, we discard the explanation sentence for each TODAY instance and only keep the additional sentence to force the model to focus only on the additional sentence.

Explanation sentence verifier. For each instance in the TODAY training set, given the same human-annotated additional sentence, we ask GPT-3 to generate three possible explanation sentences. We denote the human-annotated explanation sentence as the positive explanation and the GPT-3 generated explanation that has the lowest semantic similarity according to sentencebert as the negative sample. We finetune the base model and optimize the loss function as the negative log likelihood of the positive explanation.

[Annie: ** Todo: introduce the general algorithm **]

5 Experiment

5.1 Metrics and Settings

We measure system performance from two dimensions. We measure hard-label accuracy on TRACIE and MATRES (Zhou et al., 2021) for start-time hypotheses. We also analyze soft-label accuracy of probability change for TODAY, which is the percentage of samples where given an additional sentence (w or w/o explanation sentence) towards a certain relation direction, probability change of the final prediction in the right direction is larger than the wrong direction. [Annie: ** Is this part clear enough, maybe add a equation or sth? **]

5.2 Baselines and Systems

We use T5-large implemented by (Wolf et al., 2020) as our base temporal reasoning model. We compare our proposed models with a host of baselines, including GPT-3 and PatternTime (Zhou et al., 2021).

We compare variations of our proposed model based on the same T5-large model including T5(T), where T5 is finetuned with TRACIE training set, T5(T+O), where T5 is finetuned together with TRACIE training set and TODAY training set, T5(T+O+G), where T5 is finetuned together with TRACIE training set, TODAY training set and verifier-filterd GPT-3 generated distant supervision. We repeat this setting by replacing the TRACIE training set with MATRES training set and TRACIE + MATRES combined training set respectively. Note that we only include 1.5k (10%) training instances for MATRES to match the size of other training data. We collect 5000 initial GPT-3 generated distant supervision and 4811 remains after rule-based filter. We apply cross entropy loss for TRACIE and MATRES training set and margin ranking loss for TODAY training set and GPT-3 generated supervision.

5.3 Inference

For TODAY testing set, given the additional sentence for each instance, we utilize GPT-3 to generate three possible explanation sentences based on

the additional sentence for both relation directions of each test instance. We then rely on the explanation sentence verifier to choose the final explanation sentence, specifically we adopt the explanation sentence which has the highest score under the explanation sentence verifier.In order to enhance the explanation sentence verifier's capacity to identify an incorrect explanation sentence given a correct additional sentence, the explanation sentence verifier is especially finetuned with GPT-3 generated training set. [Annie: ** Do I need to talk in details about how it trains **]

5.4 Main Results

Table 3 shows system performances under different supervision data and loss function settings across three binary temporal benchmarks.

We observe that the average binary accuracy of TRACIE, MATRES and TODAY improves with the increasingly diversified training data and achieves a largest increase from 51.1% to 73.5% under the unified T5 training setting, which indicates that the model is being more generalized. Especially, the use of explanations contributes to an average increase of 5.6% on the average accuracy compared to merely use the temporal reasoning data, which further verifies the effeteness of explanations as guidance for models to behave correctly like a human towards this task.

We also show that the TODAY supervision contributes towards a better temporal reasoning model, with 6.7% increase on TRACIE when trained with TRACIE only, 0.5% increase on MATRES when trained with MATRES only and 6.8% increase on TRACIE and 1.5% increase on MATRES when trained together with TRACIE and MATRES. An increase of average 6% on TODAY without explanation sentence further proves that the temporal model is drifting towards the right reasoning direction to focus on the differential highlights that contribute to the temporal relation in the story.

With LLM generated distant supervision, the model performance further improves on all metrics, with an average increase of 0.45%, 0.8%, 3.83%, 1.3% on MATRES, TRACIE, TODAY and average accuracy respectively. This illustrates that LLM can provide cheap but effective distant supervision to benefit the model.

We also notice that there is a huge gap between the performance of TODAY without and with gold explanation sentence. This indicates that a correct explanation sentence can further elaborate and explain the additional sentence, i.e, the differential component. We follow the methods in 5.3 to generate an explanation for TODAY test and further improve over TODAY w/o explanation by approximate 2%, while the performance is still suboptimial compared to gold explanation sentence. The major reason is that the explanation verfier is not strong enough to choose the correct explanation from the possible two explanations of different temporal relations. We leave the research on how to generate and identify a high-quality explanation sentence for future work.

5.5 Ablation Studies and Analysis

To better understand the improvements from our models, we conduct several ablation studies. Table 4 demonstrates the results of our model with different settings of verifiers. The results have proved the effectiveness of all the verifiers. The explanation sentence verifier has the least influence. This is expected as we ask GPT-3 to generate an additional sentence followed by an explanation sentence, which largely increases its chance to be coherent as a single generation. We also utilize the rule-based filter to drop the explanations that are almost identical to the statement, which alleviates one of the major problems of GPT-3 generated explanations. The additional sentence verifier and the general verifier are more crucial as the quality of distant supervision heavily relys on if it can first correctly interpret the differences in the story and then draw a corresponding reasonable conclusion.

We also see that including more filter-verified GPT-3 data can further enhance the model performance, suggesting the usefulness of large language models to generate supervision signals to empower small models. Since the smaller T5 model with LLM distilled knowledge performs much better than the LLM itself, it also directs us to research the trade-off between model scaling and data scaling.

[Annie: ** maybe a case study of generated explanation **]

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*

Data	Loss	TRACIE	MATRES	TODAY	TODAY (gen. exp.)	TODAY (gold exp.)	Average
GPT-3	FewShot	52.3	50.1	33.8	-	-	45.4
PatternTime	Distant	77.0	73.0	54.1	59.3	67.7	68.0
T5 (O)	MR	50.6	49.8	52.9	53.7	55.7	51.1
T5 (O+G)	MR	55.4	52.3	55.0	57.8	66.5	54.2
T5 (M)	CE	52.7	81.2	52.5	55.3	57.5	62.1
T5 (M+O)	CE + MR	51.5	81.7	57.4	60.5	82.7	63.5
T5 (M+O+G)	CE + MR	49.9	82.9	61.4	61.9	**82.9**	64.8
T5 (T)	CE	66.2	63.2	52.3	55.0	56.0	60.7
T5 (T+O)	CE + MR	72.9	69.4	59.9	61.7	81.6	67.4
T5 (T+O+G)	CE + MR	73.5	68.8	62.1	63.1	82.0	68.1
T5 (M+T)	CE	66.2	82.0	52.5	54.7	58.5	66.9
T5 (M+T+O)	CE + MR	73.0	83.5	57.9	60.8	77.8	71.5
T5 (M+T+O+G)	CE + MR	73.3	83.9	**63.2**	63.1	81.6	73.5
PatternTime (all)	CE + MR	**79.9**	**86.3**	62.9	**63.4**	82.3	**76.4**

Table 3: System performances under different supervision data and loss function settings across three binary temporal benchmarks. For simplicity, we use T to denote TRACIE training data, and similarly M for MATRES, O for TODAY (ours), and G for GPT generated distant supervision. All T5 experiments are trained with the same number of steps and repeated with three seeds.We randomly select 1000 instance of MATRES for GPT-3 testing for cost control.

Data	#GPT	T	M	TODAY	Avg
Ours	1475	73.3	83.9	63.2	73.5
No Exp	1867	73.7	83.5	61.2	72.8
No Addition	2529	70.2	81.4	59.5	70.4
No General	2079	71.0	81.8	59.5	70.8
More #GPT	2483	74.6	84.0	63.2	73.9

Table 4: Ablation study for LLM generated supervision. We test the model performance under different verifier settings. We also test the setting where we include more verifier-filtered GPT-3 data (filtered by 3 verifiers).

Empirical Methods in Natural Language Processing (EMNLP), pages 5412–5417, Online. Association for Computational Linguistics.

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

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

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018a. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics.

Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018b. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Hieu Man Duc Trong, Nghia Ngo Trung, Linh Van Ngo, and Thien Huu Nguyen. 2022. Selecting optimal context sentences for event-event relation extraction. In *AAAI Conference on Artificial Intelligencel Intelligence*.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing.

In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

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

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371, Online. Association for Computational Linguistics.

A Example Appendix

This is a section in the appendix.