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 have little idea of the systems' generalizability as they rely heavily on end-to-end mechanisms. In this work, we introduce a novel task named TODAY that aims to bridge this gap and evaluates if a temporal model can make correct predictions based on the right reasons. TODAY investigates the differential analysis ability of systems, where an additional sentence is added to the context and the systems are required to tell how this subtle contextual change may affect the temporal relation of a given event pair. We also annotate human explanations of why the change affects temporal relations. We view TODAY's formulation as a close proxy to evaluating system explanations, which is the most intuitive way to sanity-check models' reasons for decision that are yet impossible to obtain automatically. We show that existing models, including GPT-3, drop to almost random guessing on TODAY, 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 (Pustejovsky et al., 2003; Chambers et al., 2014) 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

Example
Context $\mathcal{C}$: 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 1 ($\mathcal{AS}_{before}$): Dave was a scientist.
Event 1 ($\mathcal{E}_1$): Dave applied for a grant for his project. **Event 2** ($\mathcal{E}_2$): Dave worked with algae to make electricity.
Explanation: 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 of temporal differential analysis. The additional sentence $\mathcal{AS}_{before}$ has an effect on the temporal relation between the two events and shifts the temporal relation more towards Event 1 **starts before** Event 2.

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 (Zhou et al., 2021) 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 (PLMs) with annotated supervision data. Unfortunately, existing temporal benchmarks (Pustejovsky et al., 2003; Cassidy et al., 2014; Ning et al., 2018a) are flawed in two ways: 1) their training and evaluation data often come from the same distribution, and 2) they only annotate hard labels but ignores the fact that temporal labels based on common sense can be soft and nondeterministic. These two issues allow models to exploit spurious signals and annotation artifacts easily. For example, a model may learn to predict "lunch" to be before "dinner" regardless of what dates they take place, yet the majority of existing benchmarks will not challenge such beliefs. This means that the community has almost no insight into the generalizability

*Work partly done when visiting UPenn.

of current temporal reasoning models.

In this work, we bridge this gap with a novel benchmark framework that evaluates whether a temporal reasoning model is getting the correct predictions with the right reasons by properly identifying potential alternatives. Our intuition is to ask models to *explain* temporal relation predictions since the most viable way for humans to demonstrate insights into these problems is by providing satisfactory explanations. While the motivation is sound, automatically evaluating the plausibility of model explanations is extremely difficult since it is hard to obtain explanations from deep learning models nowadays. As a result, our framework employs an approximation of such explanations, which we call **temporal differential analysis**. 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 are possible if additional information in regard to the context is given. Then, we annotate a hypothetical change in the form of an additional sentence added to the beginning of the context. As Table 1 shows, this hypothetical will shift the event pair's temporal relation distribution, making it either "*more before*" or "*more after*". Each hypothetical change is also annotated with human explanations of why the change affects the temporal relation. We collect 2,241 such event pairs with a rigorous human annotation pipeline and call the resulting dataset TODAY (**t**emp**o**ral **d**ifferenti**a**l anal**y**sis). If a model is generic enough to provide proper explanations for its temporal decisions, it can also distinguish subtle context changes and understand how each change will affect the distribution of temporal relations.

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, the PatternTime model (Zhou et al., 2021) that achieves 77% binary accuracy on TRACIE (Zhou et al., 2021) - a dataset with similar contexts and events - drops dramatically to 54% on TODAY, which is barely above random guessing. 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 understanding temporal reasoning. 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. We also find that while large language models (LLMs) are not good enough for temporal differential analysis, they sometimes produce reasonable explanations for a given temporal relation. We design a pipeline that automatically collects supervision signals based on this finding. The pipeline starts with giving GPT-3 (Brown et al., 2020) an instance from TODAY and a hypothetical temporal relation and then uses GPT-3 to generate several explanations. Finally, we train an explanation verifier based on human annotation, which selects the generated explanations that are more likely to be plausible. We show that adding such explanations from GPT-3 further boosts the performance across our benchmarks.

Our contribution 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, contributes 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.

2 Related Work

Temporal Reasoning Models. Significant effort has been devoted to temporal reasoning, a challenging task that requires models to not only recognize the connection between event mentions but their context as well. Several statistical learning models (Mani et al., 2007; Ning et al., 2017, 2018b) have been proposed to characterize events based on features and learn to predict the temporal relations between event pairs. Recently, data-driven temporal reasoning approaches (Trong et al., 2022; Wang et al., 2022; Liu et al., 2021; Mathur et al., 2021; Zhou et al., 2020; Han et al., 2019) have witnessed great improvement over these feature-based models on benchmarks and are generally built upon deep neural models to predict temporal labels in an end-to-end fashion. Nevertheless, lack of interpretability has made these neural models untrustworthy to be deployed in real-world applications (Yin et al., 2022), especially in critical areas such as healthcare, finance, and government. The differential analysis approach first introduced in this paper provides a new paradigm of evaluating the interpretability of temporal reasoning models.

Temporal Relation Datasets. From different perspectives, multiple research projects have focused on the construction of temporal reasoning benchmarks. A series of remarkable datasets, TimeBank (Pustejovsky et al., 2003), TempEval 1-3 (Verhagen et al., 2007, 2010; UzZaman et al., 2013), TimeBank-Dense (Cassidy et al., 2014), RED (O'Gorman et al., 2016), MATRES (Ning et al., 2018a) and so forth, are annotated on newswire articles for events and temporal relations between events. TORQUE (Ning et al., 2020) examines models' capability in temporal reasoning in the format of reading comprehension whereas contrastive evaluation set for MATRES is introduced in (Gardner et al., 2020) to provide a local view of models' decision boundaries.

Explanations. The community has been studying explanations and how they can help the reasoning tasks such as question answering and natural language inference. Several models have been proposed (Rajani et al., 2019; Latcinnik and Berant, 2020; Kumar and Talukdar, 2020; Zhou et al., 2022), as well as evaluation benchmarks that aim to test if existing systems can properly utilize explanations (Camburu et al., 2018; Aggarwal et al., 2021). Our work is closely related to this line of effort as we attempt to build a proxy benchmark that can be automatically evaluated for temporal explanations. The recent findings on large pre-trained language models have inspired several works to use them as explanation generators (Wiegreffe et al., 2021; Marasović et al.)

3 Dataset

In this section, we introduce the evaluation framework and the collection process of TODAY.

3.1 Task overview

The TODAY dataset and its overall framework is designed to evaluate systems' ability to make temporal predictions with plausible reasons. Existing datasets, including MATRES, TORQUE, and TRACIE, annotate only common event pairs that align with human common sense. In other words, If an event pair does not strongly imply a temporal relation (e.g., over 80% confidence), it will not be annotated and tested on systems. This allows pre-trained language models with millions of parameters to exploit annotation artifacts and certain priors that do not necessarily hold in specific contexts. For example, we know "lunch" is usually before "dinner", but this also depends on if they are performed by the same subject, at the same location, and/or on the same day. Unfortunately, current models often memorize such relations as immutable facts, leading to prediction errors in instances that are less common in real life. This intuition inspires us to build a framework to evaluate exactly how much spurious information and priors models use. We later show that our framework can improve

Temporal Explanations An ideal method to evaluate if models are doing the right thing when making predictions is to let them explain why a certain prediction is made and evaluate the faithfulness and plausibility of the explanations. However, such an evaluation framework is almost impossible to achieve with current progress in natural language processing, where the two main challenges are 1) it is extremely difficult to collect gold explanations that are sufficient to cover any possible sets of explanations and 2) it is impossible to evaluate system generations using existing summarization metrics automatically.

Temporal Differential Analysis Because of the aforementioned challenges in directly evaluating system explanations, we propose an alternative that is a close proxy to the ideal form, namely temporal differential analysis. The core of temporal different analysis is to check if models can correctly identify how a subtle change to the context may affect the temporal relations of a given event pair. The intuition behind this choice is two-fold: 1) it is much easier for both annotators and models to produce an explanation if they know which dimension to focus on; 2) this provides a binary evaluation that is deterministic and trustworthy in terms of reflecting how much spurious information models are using.

Specifically, our differential analysis process is defined below. Given an original context $\mathcal{C}$, event 1 $\mathcal{E}_1$ and event 2 $\mathcal{E}_2$, we assume a gold distribution $\mathbb{D} = \{P_{before}, P_{after}, P_{same}\}$ on the temporal relation between $\mathcal{E}_1$ and $\mathcal{E}_2$ concerning $\mathcal{C}$, where $P_{before}, P_{after}, P_{same}$ are the probabilities of the temporal relation being before, after and simultaneous and they sum to 1. We then annotate two additional sentences $\mathcal{AS}_{before}$ and $\mathcal{AS}_{after}$, where the temporal relation distribution between $\mathcal{E}_1$ and $\mathcal{E}_2$ with respect to $\mathcal{AS}_{before} + \mathcal{C}$ has an increased P_{before}, while similarly the distribution using $\mathcal{AS}_{after} + \mathcal{C}$ as the context has a higher P_{after}.

Table 1 shows an example instance of our temporal differential analysis, where an additional sentence $\mathcal{AS}_{before}$ has an effect on the temporal relation between the two events and shifts the label distribution towards "before". We conduct a pilot human study for this formulation and find that it is easy to annotate and achieve substantial improvement over the explanation quality compared with directly asking for explanations on an event pair. We, therefore, adopt this formulation and create our evaluation dataset TODAY through a multi-stage annotation process as detailed below.

3.2 Dataset Construction

Following the definition of the temporal differential analysis framework above, we collect a dataset to carry out the actual evaluation. Each instance in TODAY contains a context $\mathcal{C}$, an event pair $\mathcal{E}_1$, $\mathcal{E}_2$, and an additional sentence of either $\mathcal{AS}_{before}$ or $\mathcal{AS}_{after}$. In addition, we also annotate a human explanation Exp regarding why the additional sentence affects the temporal relation between the two events. TODAY is constructed in three steps: 1) event pair generation, 2) additional sentence and explanation annotation, and 3) annotation verification and cleaning. We detail this pipeline below.

Generating $\mathcal{C}$ and $\mathcal{E}$. We randomly sample short stories from the ROCStories dataset (Mostafazadeh et al., 2016) as the context $\mathcal{C}$. For each story, we use GPT-3 to generate an implicit event phrase based on an explicit event phrase selected by GPT-3 at the same time. A sample prompt can be referred to in appendix table 5 to construct an event pair. An implicit event is a event that is not explicitly mentioned by the given context but is inferable and relevant. We do this for two main reasons: 1) events that are not explicitly mentioned by the context provide more uncertainty so that the event pair does not come with a deterministic temporal relation decided by the context; 2) this is closer to the format of TRACIE, which we aim to compare system performance changes with.

Crowdsourcing $\mathcal{AS}$ and Exp. After having $\mathcal{C}$ and $\mathcal{E}s$, we use Amazon Turk and ask crowdsourcing annotators to write potential $\mathcal{AS}_{before}$ and $\mathcal{AS}_{after}$ with respect to the provided information. The annotator is also asked to explain why they write $\mathcal{AS}$ and why it affects the temporal relation distribution in the direction that aligns with the $\mathcal{AS}$ label. We use this as Exp. We design an annotation interface that is intuitive and filled with examples, and at the same time, we require annotators to pass a rigorous qualification test to demonstrate proper understanding. We list our interface and tests in appendix figure 1 and figure 2.

Annotation Verification. We employ an additional verification stage for the human-written instances from the previous step. We provide annotators with the formatted textual entailment instance and ask if the entailment label changes in the expected direction. We collect two individual verifications per instance, and only the instances accepted by both verification annotators will appear in the test set.

3.3 Statistics

We collect 1,000 instances that are agreed upon by both verifications while constructing a silver training set with the rest 1,241 instances.

4 Modeling

In this section, we show how to make full use of TODAY's supervision signals, especially the differential explanations, to build a more generic temporal reasoning model.

However the formulation of TODAY naturally focuses on the influence of a context modification towards the temporal relation prediction while ignoring the influence of the original context itself, thus hinders its capacity for basic temporal reasoning to predict a temporal relation of an event pair given the original context only. In order to simultaneously maintain the model's basic temporal reasoning capacities and guide the model to conduct faithful reasoning by using less spurious information, we adopt a pre-trained sequence-to-sequence model as our base model and finetune the model on a temporal reasoning benchmark together with TODAY's explanation annotations. Each training instance of the model includes an instance from the temporal reasoning benchmark as an anchor to learn basic temporal reasoning and an instance from TODAY to further enhance faithful reasoning.

We adopt TRACIE's formulation (Zhou et al., 2021) to format the instance from the temporal reasoning benchmark into a textual entailment task. We compute a two-class cross-entropy loss ℓ_{CE} with logits for the instance.

Following TRACIE and for generalizability purposes, we format the TODAY instance into a textual entailment task as well, where $\mathcal{AS} + \mathcal{C}$ with Exp is the premise, and $\mathcal{E}_1$ *starts before/after* $\mathcal{E}_2$ will be the hypothesis. Here we plug in a hypothetical temporal relation in the hypothesis. If the plugged-in relation r is the same as the gold relation direction r_g where $\mathcal{AS}$ changes the relation to, the predicted entailment distribution should shift more to "entailment". Conversely, if the two relations are opposite (e.g., the plugged-in relation is "before," but the additional sentence is $\mathcal{AS}_{after}$), the system prediction is more "contradiction", we denote the opposite relation as r_o:

$$\begin{cases} p(e|(\mathcal{AS}+\mathcal{C}, Exp), r) > p(e|\mathcal{C}, r) & r = r_g \\ p(c|(\mathcal{AS}+\mathcal{C}, Exp), r) > p(c|\mathcal{C}, r) & r = r_o \end{cases} \tag{1}$$

where $p(e|pre, hyp)$ denotes the probability of being entailment given the corresponding premise and hypothesis and $p(c|pre, hyp)$ denotes the probability of being contradiction.

In order to explicitly learn from the aforementioned properties, we adopt the margin ranking loss function formulated as,

$$\begin{aligned} \ell_{MR} &= \max(0, \epsilon + p_{o_g} - p_g) \\ &\quad + \max(0, \epsilon + p_w - p_{o_w}) \\ p_g &= p(e|(\mathcal{AS}+\mathcal{C}, Exp), r_g) \\ p_{o_g} &= p(e|\mathcal{C}, r_g) \\ p_w &= p(e|(\mathcal{AS}+\mathcal{C}, Exp), r_o) \\ p_{o_w} &= p(e|\mathcal{C}, r_o) \end{aligned} \tag{2}$$

where ϵ is a margin separating the logits. The final loss function for an instance is,

$$\ell = \alpha \ell_{CE} + \ell_{MR} \tag{3}$$

where α reduces the two losses into the same scale. As a result, the proposed model is able to predict hard-label accuracy as well as soft-label probability change during inference, which leads to a more generic temporal reasoning model.

5 LLM Distant Supervision

As human explanations are inherently complicated, high-quality explanations are difficult and expensive to obtain. Recent progress in prompting LLMs provides a potentially promising alternate (Wiegreffe et al., 2022). These days, LLMs have proven to be generalist agents that work surprisingly effectively 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 differential analysis explanations from GPT-3.

5.1 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, hypothesis, additional sentence, explanation sentence) triplets followed by an unexplained (context, hypothesis) instance for which we expect the model to generate an additional sentence and explanation sentence. We demonstrate an example in table 4.

Rule-based filter. GPT-3 may cheat by generating an almost identical sentence from the context as the additional sentence or generating the exact hypothesis 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 these two categories.

5.2 General explanation verifier

Concretely, given the context C, the hypothesis with the temporal relation r, and the corresponding GPT-3 generated differential explanation of an additional sentence $\mathcal{AS}_g$ and an explanation sentence Exp_g, the general explanation verifier predicts whether the explanation is acceptable. We adopt the aforementioned model finetuned with a temporal reasoning benchmark and TODAY as the general explanation verifier. We assign the temporal relation r in the hypothesis as the gold relation direction, i.e., $r_g = r$ and the opposite relation of the temporal relation is denoted as r_o. The instances that shift more to "entailment" will be accepted, i.e.,following the annotation in 4, if:

$$\begin{aligned} &[p(e|(\mathcal{AS}_g+\mathcal{C}, Exp_g), r_g) - p(e|\mathcal{C}, r_g)] > \\ &\quad [p(e|(\mathcal{AS}_g+\mathcal{C}, Exp_g), r_o) - p(e|\mathcal{C}, r_o)], \end{aligned} \tag{4}$$

we accept the GPT-3 generated instance as a distant supervision.

5.3 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 an explanation sentence verifier to filter an inappropriate explanation sentence.

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 instance from the temporal reasoning benchmark as an input instance, we discard the explanation sentence Exp for each TODAY instance and only keep the additional sentence $\mathcal{AS}$ to force the model to focus on the additional sentence. We train the additional sentence verifier with the following loss function for the modified TODAY instance without Exp:

$$\begin{aligned} \ell^A_{MR} &= \max(0, \epsilon + p_{o_g} - p_g) \\ &\quad + \max(0, \epsilon + p_w - p_{o_w}) \\ p_g &= p(e|\mathcal{AS} + \mathcal{C}, r_g) \\ p_{o_g} &= p(e|\mathcal{C}, r_g) \\ p_w &= p(e|\mathcal{AS} + \mathcal{C}, r_o) \\ p_{o_w} &= p(e|\mathcal{C}, r_o) \end{aligned} \tag{5}$$

The final loss function for an instance in general is,

$$\ell^A = \alpha \ell_{CE} + \ell^A_{MR} \tag{6}$$

During inference, for the same instance mentioned in 5.2, it passes the additional sentence verifier only if:

$$\begin{aligned} &[p(e|(\mathcal{AS}_g + \mathcal{C}), r_g) - p(e|\mathcal{C}, r_g)] > \\ &\quad [p(e|(\mathcal{AS}_g + \mathcal{C}), r_o) - p(e|\mathcal{C}, r_o)]. \end{aligned} \tag{7}$$

Explanation sentence verifier. For each instance in the TODAY training set, given the same human-annotated additional sentence Exp, we ask GPT-3 to generate three possible explanation sentences $\{Exp_{ag}\}$. We denote the human-annotated explanation sentence Exp as the positive explanation and the GPT-3 generated explanation Exp^*_{ag} that has the lowest semantic similarity according to sentencebert as the negative explanation. We finetune the base seq-to-seq model with the positive and negative explanations and optimize the loss function as the negative log-likelihood of the positive explanation:

$$\begin{aligned} \ell^E &= -log \frac{e^{p_{pos}}}{e^{p_{pos}} + e^{p_{neg}}} \\ p_{pos} &= p(e|(\mathcal{AS} + \mathcal{C}, Exp), r_g), \\ p_{neg} &= p(e|(\mathcal{AS} + \mathcal{C}, Exp^*_{ag}), r_g), \end{aligned} \tag{8}$$

During inference, for the same instance mentioned in 5.2, it passes the explanation sentence verifier only if:

$$\begin{aligned} &p(e|(\mathcal{AS}_g + \mathcal{C}, Exp_g), r_g) > \\ &\quad p(c|(\mathcal{AS}_g + \mathcal{C}, Exp_g), r_g) \end{aligned} \tag{9}$$

Among all the GPT-3 generated instance, only the ones that passed all three verifiers will be accepted as distant supervision.

6 Experiment

6.1 Metrics and Settings

We measure system performance from two dimensions. We measure binary accuracy on TRACIE (Zhou et al., 2021) and MATRES (Ning et al., 2018a) for start-time hypotheses. We also analyze binary accuracy for TODAY to predict whether the additional sentence (w or w/o explanation sentence) makes the temporal relation shift toward the gold direction(more before/ more after) by measuring if the probability change in the gold direction is larger than the opposite direction. Given the context $\mathcal{C}$, the hypothesis with the gold/opposite temporal relation r_g/ r_o given the additional sentence $\mathcal{AS}$ w or w/o the explanation sentence Exp , following the annotation in 4, if:

$$\begin{aligned} &[p(e|(\mathcal{AS} + \mathcal{C}, Exp), r_g) - p(e|\mathcal{C}, r_g)] > \\ &\quad [p(e|(\mathcal{AS} + \mathcal{C}, Exp), r_o) - p(e|\mathcal{C}, r_o)], \end{aligned} \tag{10}$$

we label the prediction as accurate[1].

We assign α in equation 3 to be 10. All models and baselines follow a standard TE setup and default parameters. All T5 experiments are trained with the same number of steps and repeated with three seeds.

6.2 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

[1]Noted that the explanation sentence Exp will not be included in the premise if only considering the additional sentence.

Data	Loss	TRACIE	MATRES	TODAY	TODAY (gen. exp.)	TODAY (gold exp.)	Average
GPT-3	FewShot	52.3	50.1	46.8	-	-	49.7
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 2: 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-3 generated distant supervision. TODAY only includes the additional sentence. TODAY (gold exp.) includes the additional sentence and the gold explanation sentence for each instance while TODAY (gen exp.) includes the additional sentence and the explanation sentence generated by GPT-3 after filtering for each instance. Average denotes the average score of TRACIE, MATRES and TODAY for each setting. All T5 experiments are trained with the same number of steps and repeated with three seeds.

set, T5(T+O+G), where T5 is finetuned together with TRACIE training set, TODAY training set and verifier-filtered 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 remained after the 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.

6.3 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.

6.4 Main Results

Table 2 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 the largest increase from 51.1% to 73.5% under the unified T5 training setting, which indicates that the model is being more generalized. Especially if we apply all the training data to PatternTime, the average binary accuracy increases by 8.4%. 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 an explanation sentence further proves that the temporal

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 3: 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). #GPT refers to the total number of verifier-filtered GPT-3 data under each setting. T refers to TRACIE, M refers to MATRES and Avg refers to Average.

model is drifting towards the right reasoning direction to focus on the differential highlights that contribute to the temporal relation in the context.

With GPT-3 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 6.3 to generate an explanation for TODAY test and further improve over TODAY w/o explanation by approximately 2%, while the performance is still suboptimal compared to including the gold explanation sentence. The major reason is that the explanation verifier 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.

6.5 Ablation Studies and Analysis

To better understand the improvements from our models, we conduct several ablation studies. Table 3 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 hypothesis, 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 relies on if it can first correctly interpret the differences in the context 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.

7 Conclusion

We introduced a novel differential analysis framework and a dataset named TODAY that aims to interpret and evaluate if a temporal model can make correct predictions instead of using spurious information. We have demonstrated that existing temporal models fall short of the performance on TODAY. We further show that training on a temporal relation benchmark together with TODAY leads to a more generic temporal reasoning model, resulting in improved performance on TRACIE, MATRES and TODAY. Finally, we follow TODAY's formulation and distill GPT-3 to construct useful distant supervision for the model by creating a training pipeline that combines GPT-3 with weak explanation verifiers to solicit a large set of cheap and automatic explanations. Despite these advances, the gap of performance on TODAY between using additional sentence only and including human-annotated gold explanation indicates that TODAY continues to be a challenging task for future work towards generic temporal reasoning.

References

Shourya Aggarwal, Divyanshu Mandowara, Vishwajeet Agrawal, Dinesh Khandelwal, Parag Singla, and Dinesh Garg. 2021. Explanations for commonsenseqa: New dataset and models. In *Annual Meeting of the Association for Computational Linguistics*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Neural Information Processing Systems*.

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.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. 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 (EMNLP-IJCNLP)*, pages 434–444, Hong Kong, China. Association for Computational Linguistics.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Veronica Latcinnik and Jonathan Berant. 2020. Explaining question answering models through text generation. *ArXiv*, abs/2004.05569.

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.

Inderjeet Mani, Ben Wellner, Marc Verhagen, and James Pustejovsky. 2007. Three approaches to learning tlinks in timeml. *Computer Science Department*.

Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E. Peters. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*.

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.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

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

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

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

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

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

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Nazneen Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Annual Meeting of the 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*.

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.

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

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

Haoyu Wang, Hongming Zhang, Yuqian Deng, Jacob R Gardner, Muhao Chen, and Dan Roth. 2022. Extracting or guessing? improving faithfulness of event temporal relation extraction. *arXiv preprint arXiv:2210.04992*.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-AI collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 632–658, Seattle, United States. Association for Computational Linguistics.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark O. Riedl, and Yejin Choi. 2021. Reframing human-ai collaboration for generating free-text explanations. In *North American Chapter of the 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.

Fan Yin, Zhouxing Shi, Cho-Jui Hsieh, and Kai-Wei Chang. 2022. On the sensitivity and stability of model interpretations in NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2631–2647, Dublin, Ireland. 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.

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.

Ben Zhou, Kyle Richardson, Xiaodong Yu, and Dan Roth. 2022. Learning to decompose: Hypothetical question decomposition based on comparable texts. In *EMNLP*.

Let's add a sentence as the first sentence of the context to let the hypothesis more likely to hold true and explain why.
Context: 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.
Hypothesis: Test was being a good friend **starts before** he give her a really nice necklace
Add what sentence as the first sentence of the context and why is the hypothesis 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.
###
Context: 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.
Hypothesis: Test was being a good friend **starts after** he give her a really nice necklace
Add what sentence as the first sentence of the context and why is the hypothesis 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.
###
Context: 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.
Hypothesis: Tim scheduled an appointment with his dentist starts after his tooth was hurting like crazy
Add what sentence as the first sentence of the context and why is the hypothesis more likely to hold true?

Table 4: A sample prompt with an instance for two hypothetical changes to make the event pair's temporal relation "more before" or "more after".

Let's find out an event that is unmentioned but can be inferred from the story and the temporal relation between the two events are not deterministic. The new event should not be longer than ten words and include only one verb.
Context: 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.
What is an event that is unmentioned but has some role and can be inferred from the story?
Test was being a good friend
It can be inferred from She adored him for the gift.
###
Context: 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.
What is an event that is unmentioned but has some role and can be inferred from the story?
Tim scheduled an appointment with his dentist
It can be inferred from Tim's tooth was hurting like crazy.
###
Context: Lily went to a nice restaurant. She ordered a steak. To her dismay the steak was rare. Lily was rather upset. She had to send it back.
What is an event that is unmentioned but has some role and can be inferred from the story?

Table 5: A sample prompt to generate an implicit event given the context.

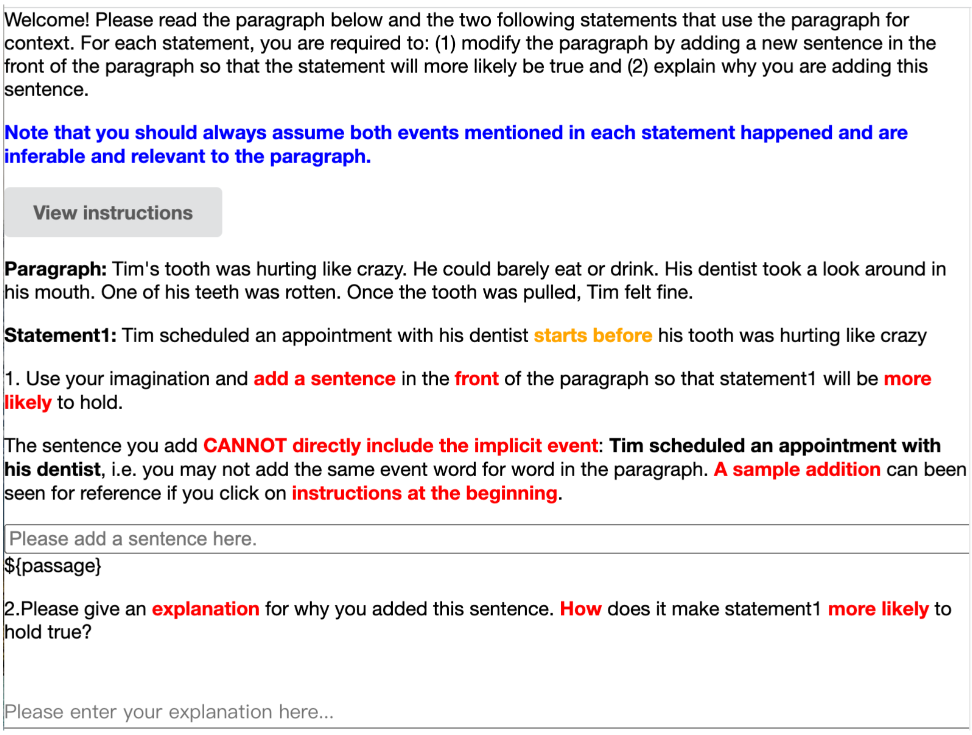

Figure 1: The interface for differential explanation annotation.

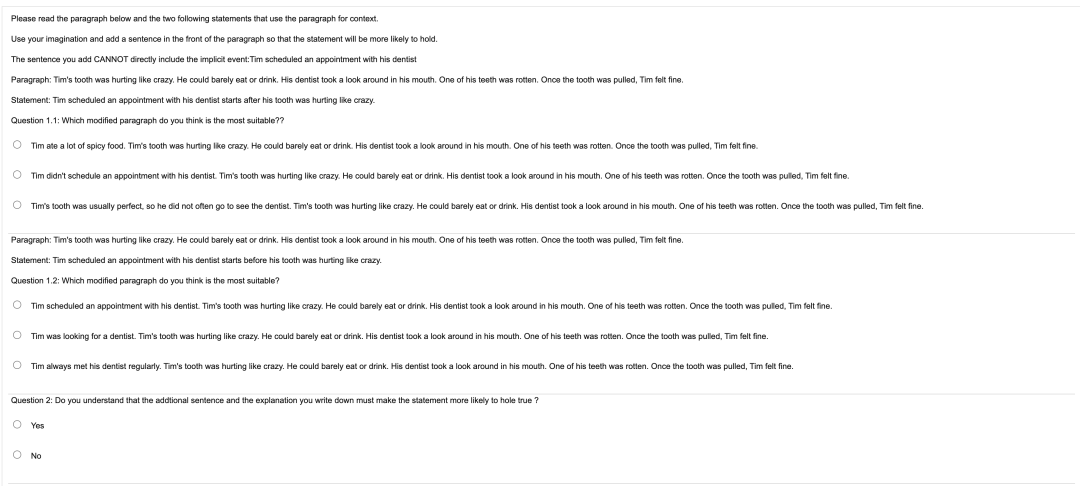

Figure 2: The interface for the qualification test of differential explanation annotation.