Extracting or Guessing? Improving Faithfulness of Event Temporal Relation Extraction

Haoyu Wang[1], **Hongming Zhang**[2], **Yuqian Deng**[1],
Jacob R. Gardner[1], **Muhao Chen**[3] **& Dan Roth**[1]
[1]Department of Computer and Information Science, UPenn
[2]Tencent AI Lab, Seattle
[3]Department of Computer Science, USC
{why16gzl, yuqiand, jacobrg, danroth}@seas.upenn.edu;
hongmzhang@tencent.com; muhaoche@usc.edu

Abstract

In this paper, we seek to improve the faithfulness of TEMPREL extraction models from two perspectives. The *first* perspective is to extract genuinely based on contextual description. To achieve this, we propose to conduct counterfactual analysis to attenuate the effects of two significant types of training biases: the *event trigger bias* and the *frequent label bias*. We also add tense information into event representations to explicitly place an emphasis on the contextual description. The *second* perspective is to provide proper uncertainty estimation and abstain from extraction when no relation is described in the text. By parameterization of Dirichlet Prior over the model-predicted categorical distribution, we improve the model estimates of the correctness likelihood and make TEMPREL predictions more selective. We also employ temperature scaling to recalibrate the model confidence measure after bias mitigation. Through experimental analysis on MATRES, MATRES-DS, and TDDiscourse, we demonstrate that our model extracts TEMPREL and timelines more faithfully compared to SOTA methods, especially under distribution shifts.

1 Introduction

Event temporal relation (TEMPREL) extraction is an essential step towards understanding narrative text, such as stories, novels, news, and guideline articles. With a robust temporal relation extractor, one can easily construct a storyline from text and capture the trend of temporally connected event mentions. TEMPREL extraction is also broadly beneficial to various downstream tasks including clinical narrative processing (Jindal and Roth, 2013; Bethard et al., 2016), question answering (Llorens et al., 2015; Meng et al., 2017; Stricker, 2021), and schema induction (Chambers and Jurafsky, 2009; Wen et al., 2021; Li et al., 2021).

Most existing TEMPREL extraction models are developed with data-driven machine learning approaches, for which recent studies also incorporate advanced learning and inference techniques such as structured prediction (Ning et al., 2017, 2018b; Han et al., 2019; Wang et al., 2020; Tan et al., 2021), graph representation (Mathur et al., 2021; Zhang et al., 2022), data augmentation (Ballesteros et al., 2020; Trong et al., 2022), and indirect supervision (Zhao et al., 2021; Zhou et al., 2021). These models are prevalently built upon pretrained language models (PLMs) and fine-tuned on a small set of annotated documents, e.g., TimeBank-Dense (Cassidy et al., 2014), MATRES (Ning et al., 2018c), and TDDiscourse (Naik et al., 2019).

A) I went to e_1:SEE the doctor. However, I was more seriously e_2:SICK. $\Longrightarrow$ e_1 AFTER e_2

B) Microsoft said it has e_3:IDENTIFIED three companies for *the China program* to run through June. The company also e_4:GIVES each participating startup in *the Seattle program* $20,000 to create software. $\Longrightarrow$ e_3 BEFORE e_4

Figure 1: Examples of unfaithful extractions. BEFORE and AFTER that follow the arrows denote the extracted TEMPREL's from the sentences by (Zhou et al., 2021).

Though these recent approaches have achieved promising evaluation results on benchmarks, whether they provide *faithful* extraction is an unexplored problem. The *faithfulness* of a relation extraction system is not simply about how much accuracy a system can offer. Instead, a faithful extractor should concern the validity and reliability of its extraction process. Specifically, when there is a TEMPREL to extract, a faithful extractor should genuinely obtain what is described in the context but not give *trivial guesses* from surface names of events or most frequent labels. Besides, when there is no relation described in the context, the system should selectively abstain from prediction.

We observe that in recent models, biases from prior knowledge in PLMs and statistically skewed training data often lead to unfaithful extractions

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 541–553

(see Fig. 1). Example A thereof exhibits a case where the model adheres to the prior knowledge where people usually **see the doctor** *after* **getting sick**, but in this context **getting sick** is obviously a consequent of **seeing the doctor**. In Example B, BEFORE is extracted due to statistical biases learned from training data that BEFORE is not only the most frequent TEMPREL between ***identify*** and ***give***, but is also the most frequent TEMPREL between the first and second event in narrative order (Gee and Grosjean, 1984). However, with a closer inspection, it can be noticed that the two events in Example B are involved in different programs, one in *the China program*, the other in *the Seattle program*. Therefore, the system should abstain from prediction and give VAGUE as output.

In this paper, we seek to improve the faithfulness of TEMPREL extraction models from two perspectives. The *first* perspective is to guide the model to genuinely extract the described TEMPREL based on a relation-mentioning context. To achieve this goal, we conduct counterfactual analysis (Niu et al., 2021) to capture and attenuate the effects of two typical types of training biases: *event bias* caused by treating event trigger names as shortcuts for TEMPREL prediction, and *label bias* that causes the model prediction to lean towards more frequent training labels. We also propose to affix tense information to event mentions to explicitly place an emphasis on the contextual description.

The *second* perspective is to teach the model to abstain from extraction when no relation is described in the text. To know when to abstain, the models need to have a good estimate of the correctness likelihood. By incorporating Dirichlet Prior (Malinin and Gales, 2018, 2019) in the training phase of current TEMPREL extraction models, we improve the predictive uncertainty estimation of the models and make the TEMPREL predictions more selective. Furthermore, since the counterfactual analysis component (from the *first* perspective) may shift the model-predicted categorical distribution, we also employ temperature scaling (Guo et al., 2017) in inference to allow for recalibrated confidence measure of the model.

The technical contributions of our work are twofolds. First, to the best of our knowledge, this is the first study on the faithfulness issue of event-centric information extraction. Evidently, the development of a faithful TEMPREL extraction system contributes to more robust and reliable machine comprehension of events and narratives. Second, we propose training and inference techniques that can be easily plugged into existing neural TEMPREL extractors and effectively improve model faithfulness by mitigating prediction shortcuts and enhancing the capability of selective prediction.

Our contributions are verified with TEMPREL extraction experiments conducted on MATRES (Ning et al., 2018c), TDDiscourse (Naik et al., 2019) and distribution-shifted version of MATRES (MATRES-DS). Particularly, we evaluate on how precise and selective our TEMPREL extraction method is on in-distribution data, and how well it generalizes under distribution shift. Experimental results demonstrate that the techniques explored within the two aforementioned perspectives bring about promising results in improving faithfulness of current models. In addition, we also apply our method to the task of timeline construction (Do et al., 2012), showing that faithful TEMPREL extraction greatly benefits the accurate construction of timelines.

2 Related Work

Event TEMPREL Extraction. Recent event TEMPREL extraction approaches are mainly built on PLMs to obtain representations of event mentions and are improved with various learning and inference methodologies. 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. (2021b); 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. Nevertheless, recent data-driven models risk amplifying bias by exacerbating biases present in the pretraining and task training data when making predictions (Zhao et al., 2017). To rectify the models' biases towards prior knowledge in PLMs and shortcuts learned from biased training examples, our work proposes several training and inference techniques, seeking to improve the faithfulness of

neural TEMPREL extractors as described in §1.

Bias Mitigation in NLP. Methods for mitigating prediction biases can be categorized as retraining and inference (Sun et al., 2019). Retraining methods address the bias in early stages or at its source. For instance, Zhang et al. (2017) masks the entities with special tokens to prevent relation extraction models from learning shortcuts from entity names, whereas several works conduct data augmentation (Park et al., 2018; Alzantot et al., 2018; Jin et al., 2020; Wu et al., 2022) or sample reweighting techniques (Lin et al., 2017; Liu et al., 2021a) to reduce biases in training. However, masking would result in the loss of semantic information and performance degradation, and it is costly to manipulate data or find proper unbiased data in temporal reasoning. Directly debiasing the training process may also hinder the model generalization on out-of-distribution (OOD) data (Wang et al., 2022). Therefore, inspired by several recent studies on debiasing text classification or entity-centric information extraction (Qian et al., 2021; Nan et al., 2021), our work adopts counterfactual inference to measure and control prediction biases based on automatically generated counterfactual examples.

Selective Prediction. Neural models have become increasingly accurate with the advances of deep learning. In the meantime, however, they should also indicate when their predictions are likely to be inaccurate in real-world scenarios. A series of recent studies have focused on resolving model miscalibration by measuring how closely the model confidences match empirical likelihoods. Among them, computationally expensive Bayesian (Gal and Ghahramani, 2016; Küppers et al., 2021) and non-Bayesian ensemble (Lakshminarayanan et al., 2017; Beluch et al., 2018) methods have been adopted to yield high quality predictive uncertainty estimates. Other methods have been proposed to use uncertainty reflected from model parameters to assess the confidence, including sharpness (Kuleshov et al., 2018) and softmax response (Hendrycks and Gimpel, 2017; Xin et al., 2021). Another class of methods adjust the models' output probability distribution by altering loss function in training via label smoothing (Szegedy et al., 2016) and Dirichlet Prior (Malinin and Gales, 2018, 2019). Besides, temperature scaling (Guo et al., 2017) also serves as a simple yet effective post-hoc calibration technique. In this paper, we model TEMPREL's with Dirichlet Prior in learning, and during inference we employ temperature scaling to recalibrate confidence measure of the model after bias mitigation.

3 Preliminaries

A document D is represented as a sequence of tokens $D = [w_1, \cdots, e_1, \cdots, e_2, \cdots, w_m]$, where some tokens belong to the set of annotated event triggers, i.e., $\mathcal{E}_D = \{e_1, e_2, \cdots, e_n\}$, and the rest are other lexemes. For a pair of events (e_i, e_j), the task of TEMPREL extraction is to predict a relation r from $\mathcal{R} \cup \{\text{VAGUE}\}$, where $\mathcal{R}$ denotes the set of TEMPREL's. An event pair is labeled VAGUE if the text does not express any determinable relation that belongs to $\mathcal{R}$. Let $\mathbf{y}^{(i,j)}$ denote the *model-predicted categorical distribution* over $\mathcal{R}$.

In order to provide a confidence estimate $\mathbf{y}$ that is as close as possible to the true probability, we first describe three separate factors (Malinin and Gales, 2018) that attribute to the predictive uncertainty for an AI system, namely epistemic uncertainty, aleatoric uncertainty, and distributional uncertainty. Epistemic uncertainty refers to the degree of uncertainty in estimating model parameters based on training data, whereas aleatoric uncertainty results from data's innate complexity. Distributional uncertainty arises when the model cannot make accurate predictions due to the lack of familiarity with the test data.

We argue that the way of handling VAGUE relations in existing TEMPREL extractors is problematic since they typically merge VAGUE into $\mathcal{R}$. In fact, VAGUE relations are complicated exception cases in the IE task, yet the annotation of such exceptions are never close to exhaustive in benchmarks, or even not given (Naik et al., 2019). In this work, we consider VAGUE relations as a source of **distributional uncertainty** and separately model them. Details are introduced in §4.2.

4 Methods

In this section, we first present how we obtain event representations and categorical distribution $\mathbf{y}$ in a local classifier for TEMPREL (§4.1). Then we introduce proposed learning and inference techniques to improve model faithfulness from the perspectives of selective prediction (§4.2) and prediction bias mitigation (§4.3), before we combine these two techniques with temperature scaling and introduce the OOD detection method in §4.4.

4.1 Local Classifier

Given that the context around an event pair (e_i, e_j)[1] has linguistic signals and temporal cues that are beneficial to TEMPREL prediction, the context of (e_i, e_j) considered in our model starts from the sentence before e_i and ends at the sentence after e_j. Inspired by Zhou and Chen (2021) for improving entity representation by prepending entity type information to entity mention spans, we add tense information of events into event trigger representations in this work. Accordingly, we enclose e_i and e_j with "@" and "#" respectively[2] and prepend their tense information to their spans with "$*$" and "$\wedge$". We provide a detailed example for affixing tense information in Appx. §A.1.

To characterize event pair (e_i, e_j), we obtain the two events' contextual representations and attention heads from PLMs. The classifer is trained to uncover the context that is critical to both events by multiplying their attentions before we send the concatenation of token embeddings and attention multiplication to a multi-layer perceptron (MLP) with $|\mathcal{R}|$ outputs. In this fashion, we obtain the $|\mathcal{R}|$-dimensional logits vector $\mathbf{z}^{(i,j)}$ and categorical distribution $\mathbf{y}^{(i,j)}$, where the probability of a label $r \in \mathcal{R}$ is given by the softmax function $\sigma(\cdot)$:

$$y_r = \sigma(\mathbf{z})_r = \frac{e^{z_r}}{\sum_{k=1}^{|\mathcal{R}|} e^{z_k}} . \tag{1}$$

4.2 Parameterization of Dirichlet Prior

As discussed in preliminaries (§3), VAGUE corresponds to complicated exception cases in inference. We model them as out-of-distribution (OOD) cases which are different from in-distribution (ID) data describing the relations in $\mathcal{R}$. The goal of providing high-quality confidence estimate $\mathbf{y}$ requires the model to yield a sharp predicted distribution centered on one of the labels in $\mathcal{R}$ when it is confident and yield a flat distribution over $\mathcal{R}$ for OOD inputs, as is shown in Fig. 2. To achieve this goal, we explicitly **parameterize a prior distribution** over categorical distributions. Because of the tractable analytic properties[3] of Dirichlet distribution (Eq. 2), we choose to parameterize a *sharp* and a *flat* Dirichlet prior over the model-predicted categorical distribution for ID and OOD inputs, respectively. The

[1]For event pair (e_i, e_j) we assume e_i is the first event in narrative order if $i < j$.

[2]Note that similar to *typed entity marker (punct)* by Zhou and Chen (2021), such enclosing has the benefit of highlighting mention spans without introducing new special tokens.

[3]$\Gamma(\cdot)$ in Eq. 2 denotes the gamma function.

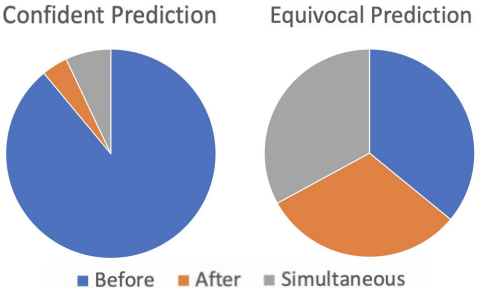

Figure 2: An illustration for desired behaviors of model predicted categorical distribution.

Dirichlet distribution is parameterized by its concentration parameters $\boldsymbol{\alpha}$, where α_0 is the precision of Dirichlet distribution. Higher values of α_0 lead to sharper, more confident predicted distributions.

$$\begin{aligned} \texttt{Dir}(\mathbf{y}; \boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{|\mathcal{R}|} \Gamma(\alpha_k)} \prod_{k=1}^{|\mathcal{R}|} y_k^{\alpha_k - 1}, \\ \alpha_k > 0,\ \alpha_0 = \sum_{k=1}^{|\mathcal{R}|} \alpha_k . \end{aligned} \tag{2}$$

To attain the aforementioned behaviors, on ID data the model is trained to minimize the KL divergence between a sharp Dirichlet distribution and the model-predicted categorical distribution:

$$\mathcal{L}_{ID} = \mathbb{E}_{\texttt{p}_{\texttt{ID}}(\boldsymbol{x})}[KL[\texttt{p}(\mathbf{y}) \parallel \texttt{Dir}(\mathbf{y}; \boldsymbol{\alpha}_s)]], \tag{3}$$

where $\texttt{p}_{\texttt{ID}}(\boldsymbol{x})$ denotes ID data and $\boldsymbol{\alpha}_s$ denotes concentration parameters of the sharp Dirichlet distribution. On OOD data, the model minimizes the KL divergence between a flat Dirichlet distribution and the model-predicted categorical distribution:

$$\mathcal{L}_{OOD} = \mathbb{E}_{\texttt{p}_{\texttt{OOD}}(\boldsymbol{x})}[KL[\texttt{p}(\mathbf{y}) \parallel \texttt{Dir}(\mathbf{y}; \boldsymbol{\alpha}_f)]], \tag{4}$$

where $\texttt{p}_{\texttt{OOD}}(\boldsymbol{x})$ denotes OOD data and $\boldsymbol{\alpha}_f$ denotes the concentration parameters of the flat Dirichlet distribution. And the total loss of the model is

$$\mathcal{L} = \lambda_1 \mathcal{L}_{ID} + \lambda_2 \mathcal{L}_{OOD}, \tag{5}$$

where the λ's are hyperparameters to balance the influence of each loss. With the parameterization of Dirichlet prior, the learning process seeks to partly enhance the model's faithfulness by outputting confident estimates when it encounters ID inputs, and outputting equivocal estimates when the context does not express any TEMPREL in the meantime.

4.3 Counterfactual Analysis

After looking into the selective prediction perspective of faithfulness, we now address the other perspective: to mitigate biases from pre-trained knowledge and the task training data during the inference

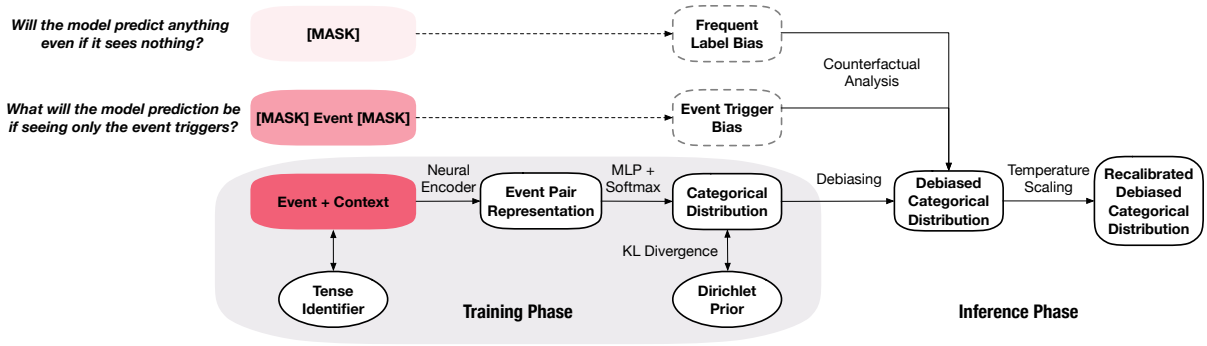

Figure 3: An overview of our approach to improving model faithfulness. In the training phase we obtain the model-predicted categorical distribution $\mathbf{y}$ with a neural encoder and parameterize a Dirichlet Prior over $\mathbf{y}$. And then we conduct counterfactual analysis to distill and mitigate biases during inference before leveraging temperature scaling to obtain recalibrated and debiased $\mathbf{y}$.

stage. Given that we have observed two types of biases in existing models, namely the event trigger bias and the frequent label bias, we ask the following questions:

- *What will the model prediction be if seeing the full context?*
- *What will the model prediction be if seeing only the event tiggers?*
- *Will the model predict anything even if it sees nothing?*

Inasmuch as we have described the learning process of the model, we know how to obtain model prediction given full context and can easily answer the first question. The second and third one, however, are hypothetical questions whose answers reflect the confounding biases that we would like to mitigate. With *attention masks* in recent PLMs (Devlin et al., 2019; Liu et al., 2019; Joshi et al., 2020; Lan et al., 2019), our model can be endowed with *imagination ability* effortlessly. By inputting a counterfactual instance with the context masked while maintaining the spans of event triggers to the model, we can obtain the model prediction given event trigger names only, which we denote by $\check{\mathbf{y}}$. And by sending an empty (counterfactual) instance, we obtain the model prediction where no textual information is given, which we denote by $\bar{\mathbf{y}}$. Intuitively, the two terms $\check{\mathbf{y}}$ and $\bar{\mathbf{y}}$ thereof provide measurements for the trigger bias and label bias.

Our goal is to use the biases assessed from model prediction (on counterfactual instances) to generate debiased categorical distribution. We remove the event trigger bias and the frequent label bias via element-wise subtraction, which is proved to be simple yet empirically effective (Qian et al., 2021):

$$\mathbf{y}' = \mathbf{y} - \beta_1 \check{\mathbf{y}} - \beta_2 \bar{\mathbf{y}}, \tag{6}$$

where $\mathbf{y}'$ denotes the debiased categorical distribution and the β's are independent parameters for balancing the terms that represent biases. We find the optimal values for β_1 and β_2 on different datasets[4] via grid beam search (Hokamp and Liu, 2017):

$$\hat{\beta}_1, \hat{\beta}_2 = \arg\max_{\beta_1, \beta_2} \psi(\beta_1, \beta_2), \quad \beta_1, \beta_2 \in [a, b], \tag{7}$$

where ψ is a metric function (e.g., F_1 scores) for evaluation, and a, b are the search boundaries.

In a nutshell, we obtain debiased categorical distribution by removing biases distilled via counterfactual inputs, thus encouraging the model to extract genuinely based on the contextual content. Nevertheless, the debiased model is not yet perfect. A minor drawback lies in that its confidence estimates might have been shifted by element-wise subtraction in Eq. 6 though it provides predictions with good evaluation results. Therefore we employ temperature scaling as our last step to allow for recalibrated confidence measure of the model.

4.4 Temperature Scaling and OOD Detection

The subtraction operation in Eq. 6 might result in negative values in $\mathbf{y}'$. To provide a proper estimate of the correctness likelihood, we first normalize the probabilities in $\mathbf{y}'$, where we replace the negative values with a small positive value and clips the values that are greater than 1:

$$\text{norm}(y'_r) = \begin{cases} \epsilon, & if\ y'_r < 0 \\ 1-\epsilon, & if\ y'_r > 1 \\ y'_r, & \text{otherwise} \end{cases} \tag{8}$$

[4]Using the development splits of datasets.

where ϵ denotes a small positive number, $r \in \mathcal{R}$. And then we use the inverse function of softmax to obtain the debiased logits vector $\mathbf{z}'$:

$$\mathbf{z}' = \sigma^{-1}(\text{norm}(\mathbf{y}')), \tag{9}$$

In this way we are able to apply temperature scaling (Guo et al., 2017) over $\mathbf{z}'$ and get the recalibrated and debiased categorical distribution $\hat{\mathbf{y}}$:

$$\hat{\mathbf{y}} = \sigma(\mathbf{z}'/T), \tag{10}$$

where $T > 0$ denotes the temperature[5].

To detect OOD inputs, we need to measure the uncertainty of the model predictions. We use the *entropy* (Eq. 11) of the final categorical distribution $\hat{\mathbf{y}}$, which captures the uncertainty encapsulated in the entire distribution. On the dev set with VAGUE examples, we find the optimal threshold of $\mathcal{H}[\hat{\mathbf{y}}]$ below which the model predictions are considered equivoques and the inputs are OOD.

$$\mathcal{H}[\hat{\mathbf{y}}] = -\sum_{k=1}^{|\mathcal{R}|} \hat{y}_k \ln(\hat{y}_k)\ . \tag{11}$$

To sum up, we improve the model faithfulness in both training and inference phase with robust event presentations, Dirichlet Prior parameterization, counterfactual analysis and temperature scaling. The entire workflow is shown in Fig. 3.

5 Experiments

In this section, we describe the experiments[6] on two tasks: TEMPREL extraction and timeline construction. We first introduce the datasets that we adopt or create for evaluation (§5.1), followed by the evaluation protocols (§5.2). Evaluation results are discussed in §5.3 before we provide a detailed ablation study and case study in §5.4 and §5.5.

5.1 Datasets

We evaluate using the following datasets, for which statistics are given in Appx. §A.4.

MATRES (Ning et al., 2018c) is a TEMPREL benchmark annotated with the multi-axis scheme that helps achieve higher inter-annotator agreements (IAA) than previous benchmark datasets (Cassidy et al., 2014; Styler et al., 2014; O'Gorman et al., 2016). Four relations are annotated for the start time comparison of event pairs in 275 documents, namely BEFORE, AFTER, SIMUTANEOUS, and VAGUE. We train our model on the training set of MATRES, and evaluate our model on the dev and test sets of MATRES, MATRES-DS and TDDiscourse, which we introduce next.

MATRES-DS is an evaluation dataset that we created with distribution shifts (DS) compared to MATRES. Since one of our goals is to mitigate the bias of event triggers in the training data, we examine whether our proposed model stays uninfluenced when the distribution of event triggers is altered. We replace frequent triggers in the MATRES dev and test sets that appear within the top 5K frequent lemmas[7] with their uncommon synonyms, and replace infrequent triggers with their frequent synonyms from the list of frequent lemmas. MATRES-DS also presents a mismatch between the training and test distributions, or dataset shift (Quinonero-Candela et al., 2008), where distributional uncertainty often arises.

TDDiscourse (Naik et al., 2019) is a dataset for discourse-level event temporal ordering, in which TEMPREL's between global long-distance event pairs are annotated. As another data source with distribution shifts compared to MATRES, we adopt the manually annotated subset of TDDiscourse, namely TDD-man, in our experiments. The TEMPREL set $\mathcal{R}_T$[8] annotated in TDDiscourse is a superset of the TEMPREL set $\mathcal{R}_M$ defined in MATRES. Given that TDD-man serves as evaluation data on which we do not train our model, a relation in $\mathcal{R}_M \cup \{\text{VAGUE}\}$ is predicted for each pair of events in the test set of TDD-man.

5.2 Evaluation Protocols

For **event TEMPREL extraction**, we compare our model with the current and previous SOTA models (Trong et al., 2022; Mathur et al., 2021) trained on MATRES. The models are evaluated on not only how precise and selective their extraction is on ID data (MATRES), but are also examined for their generalizability under distribution shifts (MATRES-DS and TDD-man). We report micro-F_1 score as an evaluation metric following previous papers. We also report macro-F_1, which reflects the fairness of model prediction, and expected calibra-

[5] T is obtained by minimizing the negative log likelihood on the dev set. We refer readers to Appx. §A.3 for details.

[6] We refer readers to Appx. §A.5 for the discussion of experimental setup.

[7] https://www.wordfrequency.info/

[8] $\mathcal{R}_T$ = {BEFORE, AFTER, SIMULTANEOUS, INCLUDES, IS INCLUDED}.

Model	MATRES micro-F_1	MATRES macro-F_1	MATRES ECE	MATRES-DS micro-F_1	MATRES-DS macro-F_1	MATRES-DS ECE	TDD-man micro-F_1	TDD-man macro-F_1	TDD-man ECE
Mathur et al. (2021)	82.3	55.7	12.8	76.7	52.3	16.3	82.1	52.8	20.3
Trong et al. (2022)	**83.4**	**56.4**	13.0	77.9	52.7	15.4	82.7	52.3	14.2
Ours	82.7	56.3	3.4	**78.7**	**54.7**	4.0	**83.1**	52.9	**5.8**
Ours w/o TI	81.8	55.2	**2.0**	77.3	52.4	8.6	79.5	**66.4**	21.9
Ours w/o DP	81.3	55.2	11.8	77.5	52.0	12.9	79.3	50.5	14.5
Ours w/o CA	80.3	54.7	5.0	78.6	52.9	**3.4**	83.0	52.7	6.4
Ours w/o TS	82.6	56.1	49.6	**78.7**	**54.7**	15.8	**83.1**	52.9	31.0

Table 1: Model performance on MATRES, MATRES-DS, and TDD-man for event TEMPREL extraction. The results of ablation study are shown in the last four rows, where TI, DP, CA and TS respectively stand for the four components in our model: Tense Information, Dirichlet Prior, Counterfactual Analysis and Temperature Scaling. Note that the numbers we report on MATRES-DS and TDD-man are model performances *under distribution shifts*.

Model	MATRES Acc	MATRES MED	MATRES-DS Acc	MATRES-DS MED	TDD-man Acc	TDD-man MED
Mathur et al. (2021)	43.5	1.44	32.1	1.75	37.3	1.49
Trong et al. (2022)	44.7	1.36	28.0	1.96	30.5	1.55
Ours	**48.2**	**1.28**	**43.5**	**1.55**	**51.7**	**1.06**
Ours w/o TI	45.8	1.37	34.5	1.66	27.1	1.87
Ours w/o DP	38.7	1.48	28.6	1.93	23.3	1.85
Ours w/o CA	43.5	1.34	39.3	1.63	49.7	1.11
Ours w/o TS	**48.2**	**1.28**	**43.5**	**1.55**	**51.7**	**1.06**

Table 2: Model performance on MATRES, MATRES-DS, and TDD-man for timeline construction. The metrics are exact match accuracy (Acc) and minimum edit distance (MED) between prediction and ground truth.

tion error (ECE) that approximates the *difference in expectation between confidence and accuracy*. The definition of ECE is provided in Appx. §A.2.

We also apply our model to the **timeline construction** task, where the goal is to sort a list of events in a document in chronological order. To construct the timeline, the model first constructs a directed graph G with predicted non-VAGUE TEMPREL's between every event pairs. Then, edges in G with lowest confidence scores are removed until G becomes a directed acyclic graph (DAG). Finally, the timeline is generated as the linear ordering of the vertices in the DAG by topological sorting. In this way, we circumvent the possible conflicts in model predictions for timeline construction and the *faithful removal* of least confident edges serves as an examination on the quality of model-predicted confidence. On the three datasets, we report the accuracy of exact match and the average minimum edit distance between predicted and ground truth timelines as evaluation metrics.

5.3 Results

In Tab. 1, we report the TEMPREL extraction results. On MATRES, the SOTA model (Trong et al., 2022) still offers the best performance in terms of micro-F_1 whereas our model achieves comparable macro-F_1 score and lower calibration error. In contrast, our proposed *faithful* TEMPREL extractor outperforms baseline methods in terms of all evaluation metrics under the dataset shifts caused by replacement of event triggers in MATRES-DS and longer context distances between global event pairs in TDD-man. Specifically, our model shows a significant gain of 2.0% macro-F_1 and 0.8% micro-F_1 over the SOTA model on MATRES-DS and surpasses the previous SOTA model on TDD-man by 1.0% micro F_1, not to mention the improvements on confidence calibration. We attribute this superior performance under dataset shifts to the mitigation of biases from prior knowledge and training set statistics as well as the techniques we employ to improve predictive uncertainty estimation. For a visual illustration of model calibration, we present the reliability diagram that plots the expected sample accuracy as a function of confidence in Fig. 4.

Tab. 2 exhibits similar observations: our model outperforms both baselines on timeline construction by a large margin in terms of both metrics. Specifically, under dataset shifts within MATRES-DS and TDD-man, our model surpasses the best baseline by 11.4% and 14.4% in accuracy, while drastically reducing the minimum edit distance by relatively 11.4% and 28.9%. Evidently, the capabilities of selective prediction and bias mitigation make our model stand out in complex scenarios like timeline construction, whereas the bias and inferior calibration of existing models exacerbate unfaithful extractions when multiple decisions have to be made simultaneously.

5.4 Ablation Study

To analyze the effect of each model component, we conduct an ablation study of our model where we

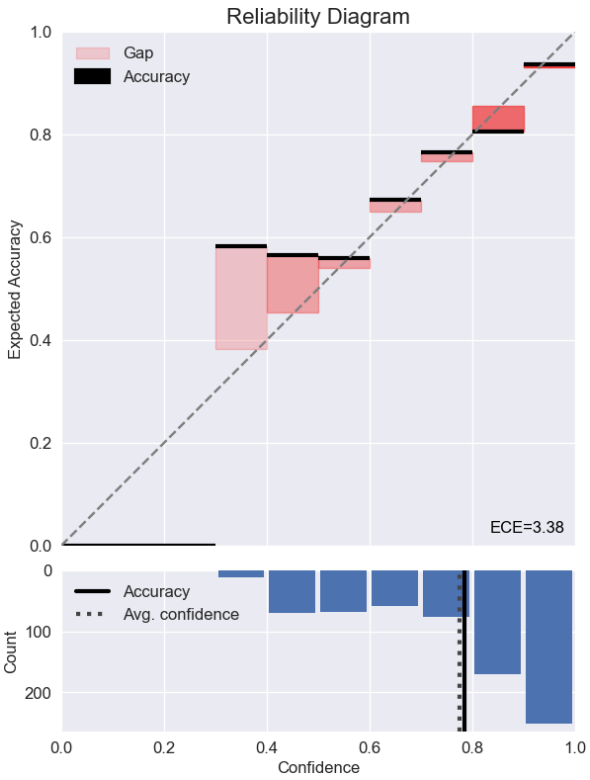

Figure 4: Reliability diagram and confidence histogram of our model predictions on test set of MATRES.

remove one component at a time (see Tab. 1)[9]. We observe on MATRES that, without the counterfactual analysis component, the model performance becomes worse by 2.4% in micro-F_1 and 1.6% in macro-F_1. Under dataset shifts, the model performance is reduced by 3.8% in micro-F_1 and 2.4% in macro-F_1 on TDD-man without the parameterization of Dirichlet Prior. The model performance in terms of F_1 scores is slightly influenced by taking away the temperature scaling component while the model calibration severely degrades.

From the ablation results in Tab. 2, we notice that temperature scaling has modest effects on the model performances, while Dirichlet Prior plays the most important role towards faithful timeline construction. It is also noteworthy that tense information considerably benefits the model to generalize well under distribution shifts in that it provides a useful feature applicable to all domains.

5.5 Case Study

As shown in Fig. 5, we provide a case study on timeline construction for three events. The reason why our model predicts AFTER for the third pair is probably due to the misleading temporal cues in text,

[9]We leverage cross-entropy as the training loss when we remove Dirichlet Prior in the training phase.

A new Essex County task force began delving Thursday into the e_1:SLAYINGS of 14 people ... officials have been e_2:CAREFUL not to draw any firm conclusions, leaving open the possibility of a serial killer ... "I haven't e_3:SEEN a pattern yet," said Patricia Hurt, the Essex County prosecutor, who created the task force on Tuesday.

Model	(e_1, e_2)	(e_2, e_3)	(e_1, e_3)	Timeline
Gold	B	B	B	(e_1, e_2, e_3)
Ours	B, 0.92	B, 0.72	A, 0.51	(e_1, e_2, e_3)
Ours w/o TI	B, 0.99	A, 0.53	B, 0.63	(e_1, e_3, e_2)
Ours w/o DP	B, 0.92	A, 0.34	B, 0.52	(e_1, e_3, e_2)
Ours w/o CA	B, 0.94	B, 0.43	B, 0.49	(e_1, e_2, e_3)
Ours w/o TS	B, 0.43	B, 0.38	A, 0.36	(e_1, e_2, e_3)

Figure 5: Case study on timeline construction for one of the documents in TDD-man. The table shows predicted TEMPREL's and confidence for three event pairs, where B stands for BEFORE and A stands for AFTER. The cells in light red and light blue are wrong predictions and relations removed in timeline construction, respectively.

Thursday and Tuesday, while the long distance between events undermines the confidence for this prediction. When our model builds a directed graph with three relations, a cycle is identified and the edge with lowest confidence is removed from the graph, and thus our model constructs the correct timeline. Without tense information, the model makes wrong prediction concerning the second event whose trigger is an adjective. And without Dirichlet Prior or temperature scaling, the model calibration becomes noticeably worse.

6 Conclusion

In this paper, we investigate on improving faithfulness for event TEMPREL extraction from two perspectives. To enhance the selectiveness of model predictions, we parameterize a Dirichlet Prior over the model-predicted categorical distribution to regularize the model to behave differently with ID and OOD data. To mitigate two types of biases from PLMs and training data, we add tense information to obtain robust event representations and conduct a counterfactual analysis to reduce the risk of carrying prediction shortcuts into inference. We also employ temperature scaling to combine the two faithful perspectives, which recalibrates the confidence measure of the model after bias mitigation. Through experimental analysis on MATRES, MATRES-DS, and TDDiscourse, we demonstrate that our model faithfully extracts event temporal relations and timelines from text, so as to generalize well under distribution shifts.

7 Acknowledgments

The authors would like to thank the anonymous ACL ARR reviewers for their insightful feedback on our work. This work was supported by Contract FA8750-19-2-1004 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. This research is also based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA Contract No. 2019-19051600006 under the BETTER Program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, the Department of Defense, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. In addition, this work is supported in part by the NSF Grant IIS 2105329, an Amazon Research Award and a Cisco Research Award.

Limitations

As the event representation introduced in our method is augmented with tense information, it potentially leads to limitations when applying to languages other than English, especially tenseless languages and languages having fewer tenses. The training of our models also requires considerable GPU resources which might produce environmental impacts, though the inference stage does not take up much computational resources.

Ethics Statement

There are no direct societal implications of this work. The proposed method attempts to provide high-quality and faithful event TEMPREL extraction and timeline construction. We believe that the intellectual merits of developing robust event-centric information extraction methods are demonstrated by this work. For any information extraction methods, real-world open source articles used to extract information may contain societal biases. Extracting event-event relations from articles with such biases may spread the bias into the acquired knowledge. Yet we believe that the proposed method can benefit various downstream NLP/NLU tasks like event prediction, task-oriented dialogue systems and risk detection.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium. Association for Computational Linguistics.

Miguel Ballesteros, Rishita Anubhai, Shuai Wang, Nima Pourdamghani, Yogarshi Vyas, Jie Ma, Parminder Bhatia, Kathleen McKeown, and Yaser Al-Onaizan. 2020. Severing the edge between before and after: Neural architectures for temporal ordering of events. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5412–5417, Online. Association for Computational Linguistics.

William H. Beluch, Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. 2018. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California. Association for Computational Linguistics.

Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 501–506, Baltimore, Maryland. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. 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 602–610, Suntec, Singapore. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Quang Do, Wei Lu, and Dan Roth. 2012. Joint inference for event timeline construction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational*

Natural Language Learning, pages 677–687, Jeju Island, Korea. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

James Paul Gee and Francois Grosjean. 1984. Empirical evidence for narrative structure. *Cognitive Science*, 8(1):59–85.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

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.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International conference on learning representation*.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025.

Prateek Jindal and Dan Roth. 2013. Using soft constraints in joint inference for clinical concept recognition. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1814, Seattle, Washington, USA. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pages 2796–2804. PMLR.

Fabian Küppers, Jan Kronenberger, Jonas Schneider, and Anselm Haselhoff. 2021. Bayesian confidence calibration for epistemic uncertainty modelling. In *Proceedings of the IEEE Intelligent Vehicles Symposium (IV)*.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare Voss. 2021. The future is not one-dimensional: Complex event schema induction by graph modeling for event prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5203–5215, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. 2021a. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hector Llorens, Nathanael Chambers, Naushad UzZaman, Nasrin Mostafazadeh, James Allen, and James Pustejovsky. 2015. SemEval-2015 task 5: QA TempEval - evaluating temporal information understanding with question answering. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 792–800, Denver, Colorado. Association for Computational Linguistics.

Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Andrey Malinin and Mark Gales. 2019. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. *Advances in Neural Information Processing Systems*, 32.

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.

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

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. TDDiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249, Stockholm, Sweden. Association for Computational Linguistics.

Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering main causalities for long-tailed information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9683–9695, Online and Punta Cana, Dominican Republic. 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, 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.

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

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

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.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445, Online. Association for Computational Linguistics.

Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2008. *Dataset shift in machine learning*. Mit Press.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2018. On the convergence of adam and beyond. In *International Conference on Learning Representations (ICLR)*.

Armand Stricker. 2021. Question answering in natural language: the special case of temporal expressions. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 184–192, Online. INCOMA Ltd.

William F Styler, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C De Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the association for computational linguistics*, 2:143–154.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2021. Extracting event temporal relations via hyperbolic geometry. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8065–8077, Online and Punta Cana, Dominican Republic. 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.

Yiwei Wang, Muhao Chen, Wenxuan Zhou, Yujun Cai, Yuxuan Liang, Dayiheng Liu, Baosong Yang, Juncheng Liu, and Bryan Hooi. 2022. Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis. In *NAACL*.

Haoyang Wen, Ying Lin, Tuan Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, Xiaodong Yu, Alexander Dong, Zhenhailong Wang, Yi Fung, Piyush Mishra, Qing Lyu, Dídac Surís, Brian Chen, Susan Windisch Brown, Martha Palmer, Chris Callison-Burch, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, and Heng Ji. 2021. RESIN: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 133–143, Online. Association for Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297.

Shuaicheng Zhang, Lifu Huang, and Qiang Ning. 2022. Extracting temporal event relation with syntactic-guided temporal graph transformer. In *NAACL*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. 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.

Wenxuan Zhou and Muhao Chen. 2021. An improved baseline for sentence-level relation extraction. *CoRR*, abs/2102.01373.

A Appendix

A.1 Example of event context affixed tense information

Original context: [CLS] For his part, Fidel Castro is the ultimate political survivor.[SEP] People have PREDICTED his demise so many times, and the US has TRIED to hasten it on several occasions.[SEP] Time and again, he endures.[SEP]

Context with affixed tense information: [CLS] For his part, Fidel Castro is the ultimate political survivor.[SEP] People have @ * *Present Perfect Simple* * PREDICTED @ his demise so many times, and the US has # ∧ *Present Perfect Simple* ∧ TRIED # to hasten it on several occasions.[SEP] Time and again, he endures.[SEP]

Figure 6: An example of the original context of event pair (PREDICTED, TRIED) and the context after affixing tense information to corresponding event spans.

A.2 Definition of ECE

Expected Calibration Error (ECE) metric (Guo et al., 2017) measures exactly the difference in expectation between confidence and accuracy. Empirically it is approximated by dividing the data into M confidence based bins, i.e., B_m (where $m \in \{1, 2, ..., M\}$) contains all datapoints i for which predicted confidence p_i lies in $(\frac{m-1}{M}, \frac{m}{M}]$. If acc($B_m$) and conf($B_m$) denotes the average accuracy and prediction confidence for the points in B_m, ECE is defined as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \Big| \text{acc}(B_m) - \text{conf}(B_m) \Big|, \quad (12)$$

where n is the number of samples. The difference between acc and conf for a given bin represents the calibration gap (red bars in reliability diagrams – e.g. Fig. 4). We use ECE as the primary empirical metric to measure model calibration.

A.3 Negative Log Likelihood

Negative log likelihood is a standard measure of a probabilistic model's quality. It is also referred to as the cross entropy loss in the context of deep learning. Given a probabilistic model $\hat{\pi}(Y|X)$ and n samples, NLL is defined as:

$$\mathcal{L} = -\sum_{i=1}^{n} \log(\hat{\pi}(y_i|\bar{X}_i)) \quad (13)$$

It is a standard result that, in expectation, NLL is minimized if and only if $\hat{\pi}(Y|X)$ recovers the ground truth conditional distribution $\pi(Y|X)$. The temperature T in temperature scaling is optimized with respect to NLL on the dev sets.

A.4 Dataset Statistics

MATRES is composed of 275 news documents and the train/dev/test split is 183/72/20 documents where 6336/6404/818 event pairs are annotated respectively. The same statistics hold for MATRES-DS since we only change the event triggers in the inputs instead of the labels. In TDDiscourse, 4,000/650/1,500 and 32609/1435/4258 TEMPREL's are annotated in the train/dev/test sets of TDD-man and TDD-Auto, respectively.

A.5 Experimental Setup and Hyperparameter Setting

In the training phase, we fine-tune the pre-trained 1024-dimensional Big Bird (Zaheer et al., 2020) to encode the context of event triggers. We obtain the tense information of event triggers with an off-the-shelf tense identifier[10]. The parameters of the model are optimized using AMSGrad (Reddi et al., 2018) with the learning rate set to 5×10^{-6}, batch size set to 20, and the training process is limited to 40 epochs on a server with Nvidia A6000 GPU. All experiments are repeated with five different random seeds and the results reported are their average. To obtain $\boldsymbol{\alpha}_s$ in Eq. 3, we smooth the target means to redistribute a small amount of probability density to the other corners of the Dirichlet. In our experiments, we set $\lambda_1 = \lambda_2 = 1$ in Eq. 5. On the dev set of TDD-man the optimal β's of the model in Eq. 6 are $\beta_1 = -0.4, \beta_2 = 0.6$, where the search bounds, a and b equal to -1 and 1.

[10] https://tense-sense-identifier.herokuapp.com/