Should We Rely on Entity Mentions for Relation Extraction? Debiasing Relation Extraction with Counterfactual Analysis

Yiwei Wang[1] **Muhao Chen**[2] **Wenxuan Zhou** [2] **Yujun Cai**[3] **Yuxuan Liang**[1]
Dayiheng Liu[4] **Baosong Yang**[4] **Juncheng Liu**[1] **Bryan Hooi**[1]
[1] National University of Singapore [2] University of Southern California
[3] Nanyang Technological University [4]Alibaba Group
wangyw_seu@foxmail.com

Abstract

Recent literature focuses on utilizing the entity information in the sentence-level relation extraction (RE), but this risks leaking superficial and spurious clues of relations. As a result, RE still suffers from unintended **entity bias**, i.e., the spurious correlation between **entity mentions (names)** and relations. Entity bias can mislead the RE models to extract the relations that do not exist in the text. To combat this issue, some previous work masks the entity mentions to prevent the RE models from over-fitting entity mentions. However, this strategy degrades the RE performance because it loses the semantic information of entities. In this paper, we propose the CoRE (**Counterfactual Analysis based Relation Extraction**) debiasing method that guides the RE models to focus on the main effects of **textual context** without losing the entity information. We first construct a causal graph for RE, which models the dependencies between variables in RE models. Then, we propose to conduct counterfactual analysis on our causal graph to distill and mitigate the entity bias, that captures the causal effects of specific entity mentions in each instance. Note that our CoRE method is model-agnostic to debias existing RE systems during inference without changing their training processes. Extensive experimental results demonstrate that our CoRE yields significant gains on both effectiveness and generalization for RE. The source code is provided at: https://github.com/vanoracai/CoRE.

1 Introduction

Sentence-level relation extraction (RE) is an important step to obtain a structural perception of unstructured text (Distiawan et al., 2019) by extracting relations between **entity mentions (names)** from the **textual context**. From human oracle, textual context should be the main source of information that determines the ground-truth relations between entities. Consider a sentence "*Mary gave birth to Jerry.*"[1]. Even if we change the entity mentions from *'Jerry'* and *'Mary'* to other people's names, the relation 'parents' still holds between the subject and object as described by the textual context *"gave birth to"*.

Recently, some work aims to utilize entity mentions for RE (Yamada et al., 2020; Zhou and Chen, 2021), which, however, leak superficial and spurious clues about the relations (Zhang et al., 2018). In our work, we observe that entity information can lead to **biased** relation prediction by misleading RE models to extract relations that do not exist in the text. Fig. 1 visualizes a relation prediction from a state-of-the-art RE model (Alt et al., 2020) (see more examples in Tab. 7). Although the context describes no relation between the highlighted entity pair, the model extracts the relation as "*countries_of_residence*". Such an erroneous result can come from the spurious correlation between entity mentions and relations, or the **entity bias** in short. For example, if the model sees the relation "*countries_of_residence*" many more times than other relations when the object entity is *Switzerland* during training, the model can associate this relation with *Switzerland* during inference even though the relation does not exist in the text.

To combat this issue, some work (Zhang et al., 2017, 2018) proposes masking entities to prevent the RE models from over-fitting entity mentions. On the other hand, some other work (Peng et al., 2020; Zhou and Chen, 2021) finds that this strategy degrades the performance of RE because it loses the semantic information of entities.

For both machines and humans, RE requires a combined understanding of textual context and entity mentions (Peng et al., 2020). Humans can avoid the entity bias and make unbiased decisions by correctly referring to the textual context that describes the relation. The underlying mechanism is

[1]We use underline and wavy line to denote subject and object respectively by default.

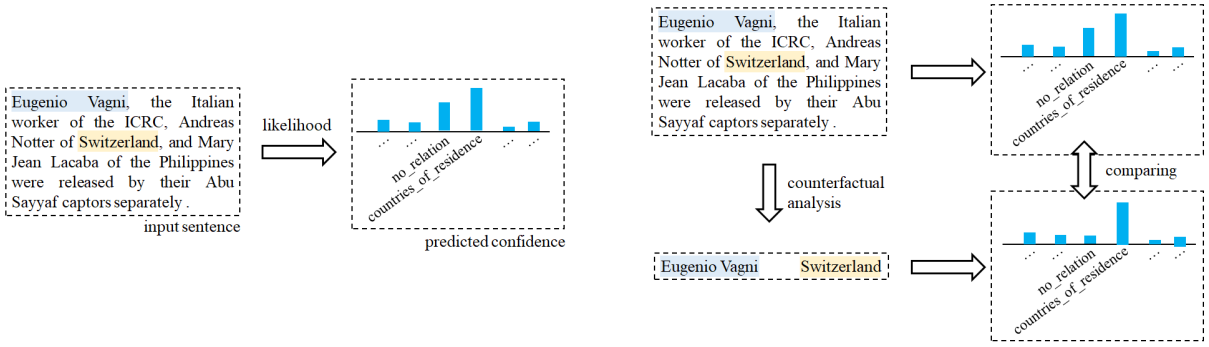

Figure 1: (*left*) An example of RE produced by LUKE (Yamada et al., 2020). In the input sentence, the subject is in blue and the object is in yellow. The ground-truth relation between the subject and object is "*no_relation*", since there is not any relation reflected by the textual context. (*right*) Our proposed counterfactual analysis for RE, which compares the original prediction (upper) with the counterfactual one (lower) to mitigate the entity bias.

causality-based (Van Hoeck et al., 2015): humans identify the relations by pursuing the main causal effect of the textual context instead of the side-effect of entity mentions. In contrast, RE models are usually *likelihood-based*: the prediction is analogous to looking up the entity mentions and textual context in a huge likelihood table, interpolated by training (Tang et al., 2020). In this paper, our idea is to teach RE models to distinguish between the effects from the textual context and entity mentions through counterfactual analysis (Pearl, 2018):

Counterfactual analysis: *If I had not seen the textual context, would I still extract the same relation?*

The counterfactual analysis essentially gifts humans the hypothesizing abilities to make decisions collectively based on the textual context and entity mentions, as well as to introspect whether the decision is deceived (see Fig. 1). Specifically, we are essentially comparing the original instance with a counterfactual instance, where only the textual context is wiped out, while keeping the entity mentions untouched. By doing so, we can focus on the main effects of the textual context without losing the entity information.

In our work, we propose a novel model-agnostic paradigm for debiasing RE, namely CoRE (Counterfactual analysis based Relation Extraction), which adopts the counterfactual analysis to mitigate the spurious influence of the entity mentions. Specifically, CoRE does not touch the training of RE models, i.e., it allows a model to be exposed to biases on the original training set. Then, we construct a causal graph for RE to analyze the dependencies between variables in RE models, which acts as a "roadmap" for capturing the causal effects of textual context and entity mentions. To rectify the test instances from the potentially biased prediction, in inference, CoRE "imagines" the counterfactual counterparts on our causal graph to distill the biases. Last but not least, CoRE performs a bias mitigation operation with adaptive weights to produce a debiased decision for RE.

We highlight that CoRE is a flexible debiasing method that is applicable to popular RE models without changing their training processes. To evaluate the effectiveness of CoRE, we perform extensive experiments on public benchmark datasets. The results demonstrate that our proposed method can significantly improve the effectiveness and generalization of the popular RE models by mitigating the biases in an entity-aware manner.

2 Related Work

Sentence-level relation extraction. Early research efforts (Nguyen and Grishman, 2015; Wang et al., 2016; Zhang et al., 2017) train RE models from scratch based on lexicon-level features. The recent RE work fine-tunes pretrained language models (PLMs; Devlin et al. 2019; Liu et al. 2019). For example, K-Adapter (Wang et al., 2020) fixes the parameters of the PLM and uses feature adapters to infuse factual and linguistic knowledge. Recent work focuses on utilizing the entity information for RE (Zhou and Chen, 2021; Yamada et al., 2020), but this leaks superficial and spurious clues about the relations (Zhang et al., 2018). Despite the biases in existing RE models, scarce work has discussed the spurious correlation between entity mentions and relations that causes such biases. Our work investigates this issue and proposes CoRE to debias RE models for higher effectiveness.

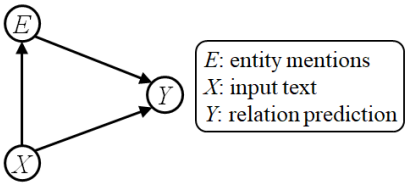

Figure 2: The causual graph of RE models.

Debiasing for Natural Language Processing. Debiasing is a fundamental problem in machine learning (Torralba and Efros, 2011). For natural language processing (NLP), some work performs data re-sampling to prevent models from capturing the unintended bias in training (Dixon et al., 2018; Geng et al., 2007; Kang et al., 2016; Rayhan et al., 2017; Nguyen et al., 2011). Alternatively, Wei and Zou (2019) and Qian et al. (2020) develop data augmentation for debiasing. Some recent work debiases the NLP models based on causal inference (Qian et al., 2021; Nan et al., 2021). In RE, how to deal with the entity bias is also an important problem. For example, PA-LSTM (Zhang et al., 2017) masks the entity mentions with special tokens to prevent RE models from over-fitting entity names, which was also adopted by C-GCN (Zhang et al., 2018) and SpanBERT (Joshi et al., 2020). However, masking entities loses the semantic information of entities and leads to performance degradation. Different from it, our CoRE model tackles entity biases based on structured causal models. In this way, we debias the RE models to focus on the textual context without losing the entity information.

3 Methodology

Sentence-level relation extraction (RE) aims to extract the relation between a pair of entities mentioned from a sentence. We propose CoRE (counterfactual analysis based Relation Extraction) as a model-agnostic technique to endow existing RE models with unbiased decisions during inference. CoRE follows the regular training process of existing work regardless of the bias from the entity mentions. During inference, CoRE post-adjusts the biased prediction according to the effects of the bias. CoRE can be flexibly incorporated into popular RE models to improve their effectiveness and generalization based on the counterfactual analysis without re-training the model.

In this section, we first formulate the existing RE models in the form of a causal graph. Then, we introduce our proposed bias distillation method to distill the entity bias with our designed counterfactual analysis. We conduct an empirical analysis to analyze how heavily the existing RE models rely on the entity mentions to make decisions. Finally, we mitigate the distilled bias from the predictions of RE models to improve their effectiveness.

3.1 Causality of Relation Extraction

In order to perform causal intervention, we first formulate the **causal graph** (Pearl et al., 2016; Pearl and Mackenzie, 2018), a.k.a., structural causal model, for the RE models as Fig. 2, which sheds light on how the textual context and entity mentions affect the RE predictions. The causal graph is a directed acyclic graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, indicating how a set of variables $\mathcal{V}$ interact with each other through the causal relations behind the data and how variables obtain their values, e.g., $(E, X) \rightarrow Y$ in Fig. 2. Before we conduct counterfactual analysis that deliberately manipulates the values of nodes and prunes the causal graph, we first revisit the conventional RE systems in the graphical view.

The causal graph in Fig. 2 is applicable to a variety of RE models and imposes no constraints on the detailed implementations. Node X is the input text. On the edge $X \rightarrow E$, we obtain the spans of subject and object entities as node E through NER or human annotations (Zhang et al., 2017). For example, in the aforementioned sentence $X =$ "*Mary gave birth to Jerry.*", the entities are $E =$ ['*Mary*', '*Jerry*'].

On the edges $(X, E) \rightarrow Y$, existing RE models take different designs. For example, C-GCN (Zhang et al., 2018) obtains the relation prediction Y by encoding entity mentions E on the pruned dependency tree of X using a graph convolutional network. IRE (Zhou and Chen, 2021) uses PLMs as the encoder for X, and marks the entity information of E with special tokens to utilize the entity information.

3.2 Bias Distillation

Based on our causal graph in Fig. 2, we diagnose how the entity bias affects inference. After training, the causal dependencies among the variables are learned in terms of the model parameters. The entity bias can mislead the models to make wrong predictions while ignoring the actual relation-describing textual context in X, i.e., biased towards the causal dependency: $E \rightarrow Y$.

The conventional biased prediction can only see the output Y of the entire graph given a sentence

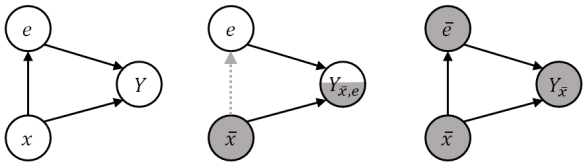

Figure 3: The original causal graph of RE models (left) together with its two counterfactual alternates for the entity bias (middle) and label bias (right). The shading indicates the mask of corresponding variables.

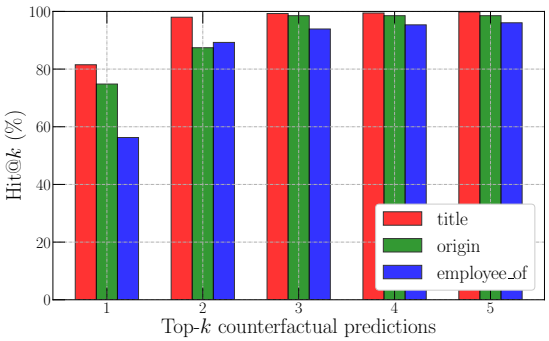

Figure 4: Hit@k (y-axis) is the fraction of the test instances, that have the original relation prediction $\arg\max_c Y_x[c]$ ranked in the top k most confident relations of the counterfactual prediction $Y_{\bar{x},e}$. We report Hit@k of the model $\text{IRE}_{\text{RoBERTa}}$ on the test instances when the original relation prediction is *title*, *employee_of*, or *origin*.

X, ignoring how specific entity mentions affect the relation prediction. However, causal inference encourages us to think out of the black box. From the graphical point of view, we are no longer required to execute the entire causal graph as a whole. In contrast, we can directly manipulate the nodes and observe the output. The above operation is termed **intervention** in causal inference, which we denote as $do(\cdot)$. It wipes out all the incoming links of a node and demands it to take a certain value.

We distill the entity bias by intervention and its induced **counterfactual**. The counterfactual means "counter to the facts", and takes one step that further assigns the hypothetical combination of values to variables. For example, we can remove the input textual context by masking X, but maintain E as the original entity mentions, as if X still exists.

We will use the input text X as our control variable where the intervention is conducted, aiming to assess its effects, due to the fact that there would not be any valid relation between entities in E if the input text X is empty. We denote the output logits Y after the intervention $X = \bar{x}$ as follows:

$$Y_{\bar{x}} = Y(do(X = \bar{x})). \quad (1)$$

Following the above notation, the original prediction Y, i.e., can be re-written as Y_x.

To distill the entity bias, we conduct the intervention $do(X = \bar{x})$ on X, while keeping the variable E as the original e, as if the original input text x had existed. Specifically, we mask the tokens in x to produce $\bar{x}$ but keep the entity mentions e as original, so that the textual context is removed and the entity information is maintained. Accordingly, the counterfactual prediction is denoted as $Y_{\bar{x},e}$ (see Fig. 3). In this case, since the model cannot see any textual context in the factual input x after the intervention $\bar{x}$, but still has access to the original entity mentions e as the inputs, the prediction $Y_{\bar{x},e}$ purely reflects the influence from e. In other words, $Y_{\bar{x},e}$ refers to the output, i.e., a probability distribution or a logit vector, where only the entity mentions are given as the input without textual context.

To investigate how heavily the state-of-the-art models rely on the entity mentions for RE, we conduct an empirical study to compare the original prediction Y_x and the counterfactual one $Y_{\bar{x},e}$. Specifically, we calculate the fraction of the test instances (y-axis) that have the original relation prediction $\arg\max_c Y_x[c]$ ranked in the top k most confident relations of the counterfactual prediction $Y_{\bar{x},e}$. This fraction is termed as Hit@k.

We present Hit@k for $\text{IRE}_{\text{RoBERTa}}$ (Zhou and Chen, 2021), a state-of-the-art RE model, in Fig. 4 on the test instances when the original relation prediction is *title*, *employee_of*, or *origin*. Higher Hit@1 means that for more instances, the model infers the same relation given only the entity mentions no matter whether the textual context is given, which imply stronger causal effects from the entity mentions $Y_{\bar{x},e}$, i.e., the models rely more heavily on the entity mentions for RE.

We observe that when $k = 1$, the Hit@1 is more than 50%, which implies that the model typically extracts the same relations even without textual context on more than a half of the instances. For a larger k, the Hit@k increases significantly and reaches more than 80% for $k \geq 2$. These observations imply a promising but embarrassing result: the state-of-the-art model relies on the entity bias for RE on many instances. The entity bias reflected by $Y_{\bar{x},e}$ can lead to the wrong extraction if the relation implied by the entity mentions does not exist in the input text. This poses a challenge to the

generalization of RE models, as validated by our experimental results (§4.3).

In addition to $Y_{\bar{x},e}$ that reflects the causal effects of entity mentions, there is another kind of bias not conditioned on the entity mentions e, but reflecting the general bias in the whole dataset, which is $Y_{\bar{x}}$. $Y_{\bar{x}}$ corresponds to the counterfactual inputs where both textual context and entity mentions are removed. In this case, since the model cannot access any information from the input after this removal, $Y_{\bar{x}}$ naturally reflects the label bias that exists in the model from the biased training. The causal graphic views of the original prediction Y_x, the counterfactual $Y_{\bar{x},e}$ for the entity bias, and $Y_{\bar{x}}$ for the label bias are visualized in Fig. 3.

3.3 Bias Mitigation

As we have discussed in §1, instead of the static likelihood that tends to be biased, the unbiased relation prediction lies in the difference between the observed outcome Y_x and its counterfactual predictions $Y_{\bar{x},e}, Y_{\bar{x}}$. The latter two are the biases that we want to mitigate from the relation prediction.

Intuitively, the unbiased prediction that we seek is the linguistic stimuli from blank to the observed textual context with specific relation descriptions, but not merely from the entity bias. The context-specific clues of the relations are key to the informative unbiased predictions, because even if the overall prediction is biased towards the relation "*schools_attended*" due to the object entity like "*Duke University*", the textual context *"work at"* indicates the relation as *"employee_of"* rather than *"schools_attended"*.

Our final goal is to use the direct effect of the textual context from X to Y for debiased prediction, mitigating (denoted as $\setminus$) the label bias and the entity bias from the prediction: $Y_x \setminus Y_{\bar{x},e} \setminus Y_{\bar{x}}$, so as to block the spread of the biases from training to inference. The debiased prediction via bias mitigation can be formulated via the conceptually simple but empirically effective element-wise subtraction operation:

$$Y_{\text{final}} = Y_x - \lambda_1 Y_{\bar{x},e} - \lambda_2 Y_{\bar{x}}, \qquad (2)$$

where λ_1 and λ_2 are two independent hyper-parameters balancing the terms for mitigating entity and label biases respectively. Note that the bias mitigation in Eq. 2 for the entity and label biases correspond to Total Direct Effect (TDE) and Total Effect (TE) in causal inference (Tang et al., 2020; VanderWeele, 2015; Pearl, 2009) respectively. We adaptively set the values of λ_1 and λ_2 for different datasets based on the grid beam search (Hokamp and Liu, 2017) in a scoped two dimensional space:

$$\lambda_1^\star, \lambda_2^\star = \arg\max_{\lambda_1,\lambda_2} \psi(\lambda_1, \lambda_2) \;\; \lambda_1, \lambda_2 \in [a, b], \quad (3)$$

where ψ is a metric function (e.g., F1 scores) for evaluation, a, b are the boundaries of the search range. We search the values of λ_1, λ_2 once on the validation set, and use the fixed values for inference on all testing instances. Since the entity types can restrict the candidate relations (Lyu and Chen, 2021), we use the entity type information, if available, to restrict the candidate relations for inference, which strengthens the effects of entity types for relation extraction.

Overall, the proposed CoRE replaces the conventional one-time prediction with Y_{final} to produce the debiased relation predictions, which essentially "thinks" twice: one for the original observation Y_x, the other for hypothesized $Y_{\bar{x}}, Y_{\bar{x},e}$.

Dataset	#Train	#Dev	#Test	#Classes
TACRED	68,124	22,631	15509	42
SemEval	6,507	1,493	2,717	19
Re-TACRED	58,465	19,584	13418	40
TACRED-Revisit	68,124	22,631	15509	42

Table 1: Statistics of datasets.

4 Experiments

In this section, we evaluate the performance of our CoRE methods when applied to RE models. We compare our methods against a variety of strong baselines on the task of sentence-level RE. Our experimental settings closely follow those of the previous work (Zhang et al., 2017; Zhou and Chen, 2021; Nan et al., 2021) to ensure a fair comparison.

4.1 Experimental Settings

Datasets. We use four widely-used RE benchmarks: TACRED (Zhang et al., 2017), SemEval (Hendrickx et al., 2019), TACRED-Revisit (Alt et al., 2020), and Re-TACRED (Stoica et al., 2021) for evaluation. TACRED contains over 106k mention pairs drawn from the yearly TAC KBP challenge. (Alt et al., 2020) relabeled the development and test sets of TACRED. Re-TACRED is a further relabeled version of TACRED after refining its label definitions. The statistics of these datasets are shown in Tab. 1.

Method	TACRED	TACRED-Revisit	Re-TACRED	SemEval
C-SGC (Wu et al., 2019)	52.1	62.8	69.8	71.3
SpanBERT (Joshi et al., 2020)	55.7	65.1	74.1	74.9
CP (Peng et al., 2020)	56.8	67.1	78.1	79.6
RECENT (Lyu and Chen, 2021)	63.3	70.5	81.1	74.6
KnowPrompt (Chen et al., 2021)	57.6	68.7	79.0	81.8
IRE_{BERT} (Zhou and Chen, 2021)	59.2	68.4	78.6	79.1
LUKE (Yamada et al., 2020)	58.8	67.5	80.2	82.1
LUKE + Resample (Burnaev et al., 2015)	59.3	68.2	80.5	82.5
LUKE + Focal (Lin et al., 2017)	59.1	67.7	80.3	82.4
LUKE + CFIE (Nan et al., 2021)	59.8	68.0	80.4	82.2
LUKE + Entity Mask (Zhang et al., 2017)	57.9	67.0	79.5	82.0
LUKE + CoRE	61.7	70.2	81.6	**83.6**
$\text{IRE}_{\text{RoBERTa}}$ (Zhou and Chen, 2021)	63.1	70.6	81.5	81.4
$\text{IRE}_{\text{RoBERTa}}$ + Resample (Burnaev et al., 2015)	63.3	71.0	81.9	81.6
$\text{IRE}_{\text{RoBERTa}}$ + Focal (Lin et al., 2017)	62.9	70.7	81.2	81.1
$\text{IRE}_{\text{RoBERTa}}$ + CFIE (Nan et al., 2021)	63.3	70.9	81.6	81.7
$\text{IRE}_{\text{RoBERTa}}$ + Entity Mask (Zhang et al., 2017)	61.4	69.3	79.6	81.2
$\text{IRE}_{\text{RoBERTa}}$ + CoRE	**64.4**	**71.8**	**82.8**	82.3

Table 2: F1-macro scores (%) of RE on the test sets of TACRED, TACRED-Revisit, Re-TACRED, and SemEval. The best results in each column are highlighted in **bold** font.

We use the widely-used F1-macro score as the main evaluation metric (Nan et al., 2021), which is the balanced harmonic mean of precision and recall, as well as F1-micro for a more comprehensive evaluation. F1-macro is more suitable than F1-micro to reflect the extent of biases, especially for the highly-skewed cases, since F1-macro is evenly influenced by the performance in each category, i.e. category-sensitive, but F1-micro simply gives equal weights to all instances (Kim et al., 2019).

Compared methods. We take the following RE models into comparison. (1) **C-SGC** (Wu et al., 2019) simplifies GCN, and combines it with LSTM, leading to improved performance over each method alone. (2) **SpanBERT** (Joshi et al., 2020) extends BERT by introducing a new pretraining objective of continuous span prediction. (3) **CP** (Peng et al., 2020) is an entity-masked contrastive pre-training framework for RE. (4) **RECENT** (Lyu and Chen, 2021) restricts the candidate relations based on the entity types. (5) **KnowPrompt** (Chen et al., 2021) is Knowledge-aware Prompt-tuning approach. (6) **LUKE** (Yamada et al., 2020) pretrains the language model on both large text corpora and knowledge graphs and further proposes an entity-aware self-attention mechanism. (7) **IRE** (Zhou and Chen, 2021) proposes an improved entity representation technique in the data preprocessing.

Among the above RE models, we apply our CoRE on LUKE and IRE. To demonstrate the effectiveness of debiased inference, we also compare with the following debiasing techniques that are applied to the same two RE models. (1) **Focal** (Lin et al., 2017) adaptively reweights the losses of different instances so as to focus on the hard ones. (2) **Resample** (Burnaev et al., 2015) up-samples rare categories by the inversed sample fraction during training. (3) **Entity Mask** (Zhang et al., 2017): masks the entity mentions with special tokens to reduce the over-fitting on entities. (4) **CFIE** (Nan et al., 2021) is also a causal inference method. In contrast to our method, CFIE strengthens the causal effects of entities by masking entity-centric information in the counterfactual predictions.

Model configuration. For the hyper-parameters of the considered baseline methods, e.g., the batch size, the number of hidden units, the optimizer, and the learning rate, we set them as suggested by their authors. For the hyper-parameters of our CoRE method, we set the search range of the hypermeters in Eq. 3 as $[-2, 2]$ and the search step 0.1. For all experiments, we report the median $F1$ scores of five runs of training using different random seeds.

4.2 Overall Performance

We implement our CoRE with LUKE and $\text{IRE}_{\text{RoBERTa}}$. Tab. 3 reports the RE results on the TACRED, TACRED-Revisit, Re-TACRED, and SemEval datasets. Our CoRE method improves the F1-macro scores of LUKE by 4.9% on TACRED, 4.0% on TACRED-Revisit, 1.7% on Re-TACRED, and 1.7 on SemEval, and improves $\text{IRE}_{\text{RoBERTa}}$

Method	TACRED	TACRED-Revisit	Re-TACRED	SemEval
LUKE (Yamada et al., 2020)	72.7	80.6	90.3	87.8
LUKE + Resample (Burnaev et al., 2015)	73.1	80.9	90.5	87.9
LUKE + Focal (Lin et al., 2017)	72.9	80.7	90.4	87.6
LUKE + CFIE (Nan et al., 2021)	73.3	80.8	90.5	88.0
LUKE + Entity Mask (Zhang et al., 2017)	72.3	80.4	90.1	87.5
LUKE + CoRE	**74.6**	**81.4**	**90.9**	**88.7**

Table 3: F1-micro scores (%) of RE on the test sets of TACRED, TACRED-Revisit, Re-TACRED, and SemEval. The best results in each column are highlighted in **bold** font.

Method	TACRED	Re-TACRED
LUKE (Yamada et al., 2020)	51.9	65.3
w/ Resample (Burnaev et al., 2015)	53.2	66.7
w/ Focal (Lin et al., 2017)	52.4	65.9
w/ CFIE (Nan et al., 2021)	52.1	65.6
w/ Entity Mask (Zhang et al., 2017)	54.5	67.1
w/ CoRE (ours)	**69.3**	**83.1**
$\text{IRE}_{\text{RoBERTa}}$ (Zhou and Chen, 2021)	56.4	68.1
w/ Resample (Burnaev et al., 2015)	58.1	70.3
w/ Focal (Lin et al., 2017)	56.8	68.7
w/ CFIE (Nan et al., 2021)	57.1	68.4
w/ Entity Mask (Zhang et al., 2017)	57.3	68.9
w/ CoRE (ours)	**73.6**	**85.4**

Table 4: F1-macro scores (%) of RE on the challenging test sets of TACRED and Re-TACRED, in which the relations implied by the entity mentions do not exist in the textual context. 'w' denotes 'with'. The best results in each column are highlighted in **bold** font.

by 1.2% on TACRED, 1.4% on TACRED-Revisit, 0.9% on Re-TACRED, and 1.8% on SemEval. As a result, our CoRE achieves substantial improvements for LUKE and $\text{IRE}_{\text{RoBERTa}}$, and enables them to outperform the baseline methods. Additionally, we report the experimental results in terms of F1-micro scores in Tab. 3, showing the improvement from CoRE on LUKE by 2.6% on TACRED, 1.0% on TACRED-Revisit, 0.7% on Re-TACRED, and 1.0% on SemEval. Overall, our CoRE method improves the effectiveness of RE significantly in terms of both F1-macro and F1-micro scores. The above experimental results validate the effectiveness and generalization of our proposed method.

Among the baseline debiasing methods, Resample, Focal, CFIE cannot distill the entity bias in an entity-aware manner like ours. Entity Mask leads to the loss of information, while our CoRE enables RE models to focus on the main effects of textual context without losing the entity information. The superiority of CoRE highlights the importance of the causal inference based entity bias analysis for debiasing RE, which compares traditional likelihood-based predictions and hypothesized counterfactual ones to produce debiased predictions. Besides, the proposed CoRE works in inference and thus can be employed on the previous already-trained models. In this way, CoRE serves as a model-agnostic approach to enhance RE models without changing their training process.

4.3 Analysis on Entity Bias

Some work argues that RE models may rely on the entity mentions to make relation predictions instead of the textual context (Zhang et al., 2018; Joshi et al., 2020). The empirical results in Fig. 3 validates this argument. Regardless of whether the textual context exists or not, the baseline RE model makes the same predictions given only entity mentions on many instances. The entity bias can mislead the RE models to make wrong predictions when the relation implied by the entity mentions does not exist in the textual context.

To evaluate whether RE models can generalize well to particularly challenging instances where relations implied by the entity mentions do not exist in the textual context, we propose a filtered evaluation setting, where we keep the test instances having the entity bias different from their ground-truth relations. In this setting, RE models cannot overly rely on the entity mentions for RE, since the entity mentions no longer provide the superficial and spurious clues for the ground-truth relations.

We present the evaluation results on the filtered test set in Tab. 4. Our CoRE method consistently and substantially improves the effectiveness of LUKE and IRE on the filtered test set and outperforms the baseline methods by a significant margin, which validates the effectiveness and generalization of our method to mitigate the entity bias in the challenging cases.

4.4 Evaluation on Fairness

According to Sweeney and Najafian (2019), the more imbalanced/skewed a prediction produced by a trained model is, the more unfair opportunities it

Method	TACRED	TACRED-Revisit	Re-TACRED	SemEval
$\text{IRE}_{\text{RoBERTa}}$ (Zhou and Chen, 2021)	61.2	59.3	57.5	54.1
$\text{IRE}_{\text{RoBERTa}}$ + Resample (Burnaev et al., 2015)	60.5	58.4	56.8	53.5
$\text{IRE}_{\text{RoBERTa}}$ + Focal (Lin et al., 2017)	60.9	58.9	57.1	53.7
$\text{IRE}_{\text{RoBERTa}}$ + CFIE (Nan et al., 2021)	60.1	57.8	56.2	52.9
$\text{IRE}_{\text{RoBERTa}}$ + Entity Mask (Zhang et al., 2017)	61.5	60.1	57.3	54.2
$\text{IRE}_{\text{RoBERTa}}$ + CoRE	**57.3**	**55.6**	**54.3**	**50.8**

Table 5: Experimental results (unfairness; %) of Relation Extraction on the test sets of TACRED, TACRED-Revisit, Re-TACRED, and SemEval (lower is better). The best results in each column are highlighted in **bold** font.

LUKE + CoRE	**61.7**	Δ	**IRE + CoRE**	**64.4**	Δ
w/o CoRE	58.8	2.9↓	w/o CoRE	63.1	1.3↓
w/o EBM	59.5	2.2↓	w/o EBM	63.4	1.0↓
w/o LBM	60.8	0.9↓	w/o LBM	63.9	0.5↓
w/o BSH	60.1	1.6 ↓	w/o BSH	63.8	0.6↓

Table 6: Ablation study based on the TACRED dataset. The analyzed model components include entity bias mitigation operation (EBM), the label bias mitigation operation (LBM) and the beam search for hyper-parameters (BSH). 'w/o' denotes 'without'. ↓ denotes performance drop in terms of F1-macro scores.

gives over predefined categories, and the more unfairly discriminative the trained model is. We thus follow previous work (Xiang et al., 2020; Sweeney and Najafian, 2019; Qian et al., 2021) to use the metric – *imbalance divergence* – to evaluate how imbalanced/skewed/unfair a prediction P is :

$$D(P, U) = JS(P \| U), \tag{4}$$

where $D(\cdot)$ is defined as the distance between P and the uniform distribution U. Specifically, we use the *JS* divergence as the distance metric since it is symmetric (i.e., $JS(P\|U) = JS(U\|P)$) and strictly scoped (Fuglede and Topsoe, 2004). Based on this, to evaluate the entity bias of a trained RE model, we average the following *relative entity mention imbalance* (REI) measure over all the testing instances containing whichever entity mentions:

$$\text{REI} = \frac{1}{\mathcal{E}} \sum_{e \in \mathcal{E}} D(P(\{x | e \in x \wedge x \in \mathcal{D}\}), U), \tag{5}$$

where x is an input instance, $\mathcal{D}$ is the testing set, $P(x)$ is the prediction output, e is an entity mention, and $\mathcal{E}$ is the corpus of entity mentions. This metric captures the distance between all predictions and the fair uniform distribution U.

We follow the experimental settings in §4.2 and report the fairness test in Tab. 5. The results show that our CoRE method reduces the imbalance metrics (lower is better) when employed on $\text{IRE}_{\text{RoBERTa}}$ significantly and consistently, indicating that it is helpful to mitigate the entity bias.

4.5 Ablation and Case Study

We conduct ablation studies on CoRE to empirically examine the contribution of its main technical components. including the entity bias mitigation operation (EBM), the label bias mitigation operation (LBM) and the beam search for hyper-parameters (BSH).

We report the experimental results of the ablation study in Tab. 6. We observe that removing our CoRE causes serious performance degradation. This provides evidence that using our counterfactual framework for RE can explicitly mitigate biases to generalize better on unseen examples. Moreover, we observe that mitigating the two types of biases is consistently helpful for RE. The key reason is that the distilled label bias provides an instance-agnostic offset and the distilled entity bias provides an entity-aware one in the prediction space, which makes the RE models focus on extracting relations on the textual context without losing the entity information. Meanwhile, the beam search for hyper-parameters effectively finds two dynamic scaling factors to amplify or shrink two biases, making the biases be mitigated properly and adaptively.

Tab. 7 gives a qualitative comparison example between CoRE and $\text{IRE}_{\text{RoBERTa}}$ on TACRED. The results show that the state-of-the-art RE model $\text{IRE}_{\text{RoBERTa}}$ returns the relations that do not exist in the textual context between the considered entities. For example, given "*Bibi drew the ire of fellow farmhands after a dispute in June 2009, when they refused to drink water she collected and she refused their demands that she convert to Islam.*", there is no relation between *Bibi* and *Islam* exists in the text but the baseline model believes that the relation between them is "*religion*". The counterfactual prediction can account for this disappointing result,

Input sentence	Original	Debiased	Counterfactual
More than 1,100 miles (1,770 kilometers) away, Alan Gross passes his days in a Cuban military hospital, watching baseball on a small television or jamming with his jailers on a stringed instrument they gave him.	origin ✗	countries_of_residence ✓	origin
He said that according to his investigation, Bibi drew the ire of fellow farmhands after a dispute in June 2009, when they refused to drink water she collected and she refused their demands that she convert to Islam.	religion ✗	no_relation ✓	religion
ShopperTrak also estimates foot traffic in the U.S. was 11.2 percent below what it would have been Sunday if the blizzard had not occurred and 13.9 percent below what it could have been Monday.	country_of_headquarters ✗	no_relation ✓	country_of_headquarters

Table 7: A case study for $\text{IRE}_{\text{RoBERTa}}$ and our CoRE on the relation extraction dataset TACRED. Underlines and wavy lines highlight the subject and object entities respectively. We report the original prediction, the corresponding counterfactual prediction and the debiased prediction.

where given only the entity mentions *Bibi* and *Islam*, the RE model returns the relation "*religion*" without any textual context. This implies that the model makes the prediction for the original input relying on the entity mentions, which leads to the wrong RE prediction. Our CoRE method distills the biases through counterfactual predictions and mitigates the biases to distinguish the main effects from the textual context, which leads to the correct predictions as shown in Tab. 7.

Last but not least, we conduct experiments on the fairness of different models, and present respective results in the appendix.

5 Conclusion

We have designed a counterfactual analysis based method named CoRE to debias RE. We distill the entity bias and mitigate the distilled biases with the help of our causal graph for RE, which is a road map for analyzing the RE models. Based on the counterfactual analysis, we can analyze the side-effects of entity mentions in the RE and debias the models in an entity-aware manner. Extensive experiments demonstrate that our methods can improve the effectiveness and generalization of RE. Future work includes analyzing the effects of other factors that can cause bias in natural language processing.

Acknowledgement

The authors would like to thank the anonymous reviewers for their discussion and feedback.

Muhao Chen and Wenxuan Zhou are supported by the National Science Foundation of United States Grant IIS 2105329, and by the DARPA MCS program under Contract No. N660011924033 with the United States Office Of Naval Research. Except for Muhao Chen and Wenxuan Zhou, this paper is supported by NUS ODPRT Grant R252-000-A81-133 and Singapore Ministry of Education Academic Research Fund Tier 3 under MOEs official grant number MOE2017-T3-1-007.

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.

Evgeny Burnaev, Pavel Erofeev, and Artem Papanov. 2015. Influence of resampling on accuracy of imbalanced classification. In *Eighth international conference on machine vision (ICMV 2015)*, volume 9875, page 987521. International Society for Optics and Photonics.

Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2021. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction.

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.

Bayu Distiawan, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.

Guang-Gang Geng, Chun-Heng Wang, Qiu-Dan Li, Lei Xu, and Xiao-Bo Jin. 2007. Boosting the performance of web spam detection with ensemble under-sampling classification. In *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, volume 4, pages 583–587. IEEE.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2019. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *arXiv preprint arXiv:1911.10422*.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. *arXiv preprint arXiv:1704.07138*.

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.

Qi Kang, XiaoShuang Chen, SiSi Li, and MengChu Zhou. 2016. A noise-filtered under-sampling scheme for imbalanced classification. *IEEE transactions on cybernetics*, 47(12):4263–4274.

Kang-Min Kim, Yeachan Kim, Jungho Lee, Ji-Min Lee, and SangKeun Lee. 2019. From small-scale to large-scale text classification. In *The World Wide Web Conference*, pages 853–862.

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.

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*.

Shengfei Lyu and Huanhuan Chen. 2021. Relation classification with entity type restriction. *arXiv preprint arXiv:2105.08393*.

Guoshun Nan, Jiaqi Zeng, Rui Qiao, Zhijiang Guo, and Wei Lu. 2021. Uncovering main causalities for long-tailed information extraction. *arXiv preprint arXiv:2109.05213*.

Hien M Nguyen, Eric W Cooper, and Katsuari Kamei. 2011. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1):4–21.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 39–48, Denver, Colorado. Association for Computational Linguistics.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Judea Pearl. 2018. Causal and counterfactual inference. *The Handbook of Rationality*, pages 1–41.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from context or names? an empirical study on neural relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672.

Chen Qian, Fuli Feng, Lijie Wen, Li Lin, and Tat-Seng Chua. 2020. Enhancing text classification via discovering additional semantic clues from logograms. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1201–1210.

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.

Farshid Rayhan, Sajid Ahmed, Asif Mahbub, Rafsan Jani, Swakkhar Shatabda, and Dewan Md Farid. 2017. Cusboost: Cluster-based under-sampling with boosting for imbalanced classification. In *2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS)*, pages 1–5. IEEE.

George Stoica, Emmanouil Antonios Platanios, and Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13843–13850.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725.

Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE.

Nicole Van Hoeck, Patrick D Watson, and Aron K Barbey. 2015. Cognitive neuroscience of human counterfactual reasoning. *Frontiers in human neuroscience*, 9:420.

Tyler VanderWeele. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307, Berlin, Germany. Association for Computational Linguistics.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. 2019. Simplifying graph convolutional networks. In *International conference on machine learning*, pages 6861–6871. PMLR.

Liuyu Xiang, Guiguang Ding, and Jungong Han. 2020. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, pages 247–263. Springer.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium. Association for Computational Linguistics.

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.

Wenxuan Zhou and Muhao Chen. 2021. An improved baseline for sentence-level relation extraction. *arXiv preprint arXiv:2102.01373*.