Towards Model Robustness: Generating Contextual Counterfactuals for Entities in Relation Extraction

Mi Zhang
School of Computer Science, Wuhan University,
Wuhan, Hubei, China
mizhanggd@whu.edu.cn

Tieyun Qian*
School of Computer Science, Wuhan University,
Wuhan, Hubei, China
qty@whu.edu.cn

Ting Zhang
School of Computer Science, Wuhan University,
Wuhan, Hubei, China
tingzhang_17@whu.edu.cn

Xin Miao
School of Computer Science, Wuhan University,
Wuhan, Hubei, China
miaoxin@whu.edu.cn

ABSTRACT

The goal of relation extraction (RE) is to extract the semantic relations between/among entities in the text. As a fundamental task in information systems, it is crucial to ensure the robustness of RE models. Despite the high accuracy current deep neural models have achieved in RE tasks, they are easily affected by spurious correlations. One solution to this problem is to train the model with counterfactually augmented data (CAD) such that it can learn the causation rather than the confounding. However, no attempt has been made on generating counterfactuals for RE tasks.

In this paper, we *formulate the problem of automatically generating CAD for RE tasks* from an entity-centric viewpoint, and *develop a novel approach to derive contextual counterfactuals* for entities. Specifically, we exploit two elementary topological properties, i.e., the centrality and the shortest path, in syntactic and semantic dependency graphs, to first identify and then intervene on the contextual causal features for entities. We conduct a comprehensive evaluation on four RE datasets by combining our proposed approach with a variety of RE backbones. Results prove that our approach not only improves the performance of the backbones but also makes them more robust in the out-of-domain test [1].

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

relation extraction, counterfactual reasoning, semantic and syntactic graph topology

ACM Reference Format:
Mi Zhang, Tieyun Qian*, Ting Zhang, and Xin Miao. 2023. Towards Model Robustness: Generating Contextual Counterfactuals for Entities in Relation Extraction. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543507.3583504

*Corresponding Author
[1]Our code and data are available at https://github.com/NLPWM-WHU/CoCo.

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583504

1 INTRODUCTION

Relation extraction (RE) aims to extract the semantic relations between/among entities in the text. It serves as a fundamental task in information systems, and facilitates a range of downstream applications such as knowledge graph construction and question answering. Deep neural models have made substantial progress in many research fields including RE. However, existing studies [7, 40] have shown that neural models are prone to be unstable due to the spurious correlations. For example, given two classes '*place-of-birth*' and '*date-of-birth*' in RE, if there are many examples like 'Newton was born in 1643' and 'Einstein was born in 1879' and a few examples like 'Newton was born in England' in training data, the model will treat '*born in*' as a spurious feature of the label '*date-of-birth*', and assigns this wrong label to 'Biden was born in Pennsylvania' in test data. The robustness and generalization of the model can be severely affected by spurious correlations. Consequently, it is desirable to train robust models by generating counterfactually augmented data (CAD) [9] such that the model can distinguish causal and spurious patterns.

There has been an increasing interest in generating counterfactuals in information systems, including sentiment classification [1, 28, 34], counterfactual explanations [16], and dialogue generation [42]. Early research often employs human annotators and designs human-in-the-loop systems [9, 23]. Most of recent studies generate counterfactuals with semantic interventions using templates, lexical and paraphrase changes, and text generation methods [4, 13, 21, 26]. A few of them incorporate the syntax into language models [36].

The aforementioned methods are all proposed for sentence level coarse-grained tasks. The key challenge in the fine-grained RE task is that it involves two or more entities which should remain unchanged during the intervention, otherwise the problem itself also changes. Existing semantic intervention methods tend to select content words for generating counterfactuals. The reason is that they follow "the minimal change" principle [9, 34], where the content words like entities have a large probability to be replaced since they convey more information than function words. This is not desirable in our task. The syntactic intervention method [36] replaces each

type of dependency relation between two words with a randomized one. Such a method cannot produce counterfactuals that flip the label yet keeping entities unchanged for our RE task, either.

In view of this, we *introduce the problem of automatically generating CAD into RE tasks for the first time*. To meet the condition of invariant entities, we formulate it from an entity-centric viewpoint. We then develop a novel approach to derive counterfactuals for the contexts of entities. Instead of directly manipulating the raw text, we deploy semantic dependency graph (SemDG) and syntactic dependency graph (SynDG) as they contain abundant information. We *exploit two elementary topological properties to identify contextual causal features for entities*. In particular, the centrality measures the importance of the word in the SynDG, which helps us recognize structurally similar entities in two samples. Meanwhile, the shortest path between two entities in SemDG captures the basic relation between them. We then generate CAD by intervening on the contextual words around one specific entity and those along the shortest path between two entities.

The contributions of this study are summarized as follows.

- To the best of our knowledge, we are the first to investigate the problem of automatical generation of CAD for the RE task, which can improve the model robustness and is important for real applications.
- We propose a novel approach which exploits the topological structures in both the semantic and syntactic dependency graphs to generate human-like counterfactuals for each original sample.
- Extensive experiments on four benchmark datasets prove that our approach significantly outperforms the state-of-the-art baselines. It is also effective in alleviating spurious associations and improving the model robustness.

2 RELATED WORK

2.1 Relation Extraction

Deep learning models have been successfully employed in RE tasks, either for extracting better semantic features from word sequences [38, 41], or incorporating syntactic features over the dependency graph [5, 14]. More recently, pre-trained language model (PLM) based methods have become the mainstream [2, 19, 32].

Despite the remarkable performance deep neural models have achieved in RE, their applications still faces big challenges. One particular concern is that these models might learn unexpected behaviors that are associated with spurious patterns.

2.2 Counterfactual Reasoning

There has been a growing line of research to learn causal associations using causal inference. Early work attempts to achieve model robustness with the help of human-in-the-loop systems to generate counterfactual augmented data [9, 23]. Recently, automatically generating counterfactuals has received more and more attention. For example, Wang and Culotta [28], Yang et al. [34] identify causal features and generate counterfactuals by substituting them with other words for sentiment analysis and text classification. Yu et al. [36] generate counterfactually examples by randomly replacing syntactic features to implicitly force the networks to learn semantics and syntax. There are aslo some methods deploy PLMs to obtain a universal counterfactual generator for texts [4, 13, 21, 24, 30].

Overall, the problem of automatical generation of CAD has not been explored in RE tasks. A seminal work [11] proposes an element intervention method for open relation extraction. However, it is designed to resolving spurious correlations rather than the generation of CAD. Our work makes the first attempt on it. Furthermore, our method takes advantage of both syntax and semantic information in dependency graphs, which allows us to generate grammatically correct and semantically readable counterfactuals. Note that our method cannot be directly applied to document level RE tasks unless a document level semantic or syntactic dependency tree/graph is constructed in advance.

3 PRELIMINARY

3.1 Task Definition

3.1.1 Relation Extraction (RE). Relation extraction tasks are mainly categorized into three types: sentence-level RE, cross-sentence n-ary RE, and document-level RE. Our proposed model is targeted for the first two types and can be extended to the last one which we leave for the future work. Formally, let $\mathcal{X} = [t_1, ..., t_m]$ be a text consisting of sentence(s) with m tokens and two or three entity mentions. $\mathcal{Y} = \{y_1, ..., y_{p-1}, y_p\}$ (y_p = None) is a predefined relation set. The RE task can be formulated as a classification problem of determining whether a relation holds for entity mentions.

3.1.2 Generating Counterfactuals in RE. Given a RE dataset $\{(X,Y)\}$ and the model $f: X \rightarrow Y$, we aim to generate a set of counterfactuals $\{(X_{cf}, Y_{cf})\}$ $(Y_{cf} \neq Y)$ [2] without changing entities. For the purpose of training a robust model, it is desirable to make X_{cf} as similar to X as possible [34], i.e., X_{cf} is syntax-preserving and semantics-reasonable.

3.2 Structural Causal Model

We introduce the structure causal model (SCM) [8] to investigate the causal relationship between data and the RE model, where random variables are vertices and an edge denotes the direct causation between two variables. Before start, we first propose our *causal questions to guide the generation of SCM*. (1) What would happen if the important syntactic structure *SY* around the entity is changed? (2) What would happen if the semantic path *SE* between two entities in the text changes?

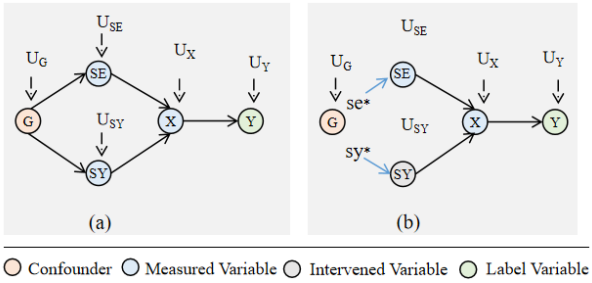


Figure 1: An illustration of structural causal models describing the mechanism of the causal inference for RE models.

[2]Note that several studies treat the label-preserved samples as counterfactuals though they only involve spurious features and can be used as ordinary augmented data. For clarity, we denote the label-flipped samples as counterfactuals since they reveal the causation between the causal feature and the label.

Based on the causal questions, we consider two important factors SY and SE as variables in SCM to capture the causal relation in the text. As shown in Fig. 1 (a), there is a confounding variable G that influences the generation of SY and SE. Y is the label variable, and U_* represents the unmeasured variable. $X \rightarrow Y$ means there exits a direct effect from X to Y. Furthermore, the path $SY/SE \rightarrow X \rightarrow Y$ denotes SY/SE has an indirect effect on Y via a mediator X. Y can be calculated from the values of its ancestor nodes, which is formulated as: $Y_{sy/se,x} = f(SY = sy/SE = se, X = x)$, where $f(.)$ is the value function of Y.

To estimate causal effects of the variable SY/SE on the target variable Y, we need to block the influence of the confounding variable G on SY/SE, and see how Y is changed by a unit intervention when fixing the value of sy/se as sy^*/se^*, as shown in Fig. 1 (b).

4 MODEL OVERVIEW

This section presents our approach for automatically generating counterfactuals by substituting causal compositions with candidate ones to enhance the robustness of RE models. One of our key insights is adopting the graph formulation to incorporate the rich information in syntactic dependency graph (SynDG) and semantic dependency graph (SemDG).

SynDG pays attention to the role of *non-substantive words* such as prepositions in the sentence. Existing work [5] has shown the effectiveness by applying SynDGs to RE models with graph convolutional networks. We also exploit SynDG but our goal is to identify causal syntactic features. Moreover, we introduce SemDG into the field of causality analysis. Our intuition is that SemDG reveals the semantic relationship between *substantive words*, and crosses the constraints of the surface syntactic structure. In other words, SemDG provides complementary information for SynDG.

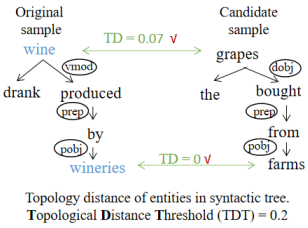
A counterfactual, denoted as X_{cf}, is a sample which has the most similar semantic or syntactic structure with the original sample X_{ori} but has a different label. Note that we cannot change the entities during the interventions for samples in RE tasks. Hence we propose our **entity-centric framework** to generate counterfactuals by first identifying and then intervening on contextual causal features for entities via topological based analysis in SynDG and SemDG. The identification of causal features consists of two steps.

- To identify **the syntactic causal composition around the entity**, we conduct the centrality analysis for entities in two samples since centrality measures how importance a node is in the graph.
- To identify **the semantic causal composition between two entities**, we employ the shortest dependency path (SDP) between them since SDP retains the most relevant information while eliminating irrelevant words (noises) in the sentence [31].

4.1 Generating Syntactic Counterfactuals

Since a counterfactual should flip the label of the original sample X_{ori}, the candidate substitute X_{can} which will be used for identifying causal features is randomly chosen from the training samples with different class labels [3]. Moreover, since entities are of the most importance in RE tasks, the entities in the substitute should be most similar with those in X_{ori}. In view of this, we first identify the syntactically similar entity nodes in SynDG using the centrality metric, which is proposed to account for the importance of nodes in a graph (network). We employ three types of centralities including betweenness centrality (BC), closeness centrality (CC), and degree centrality (DC) for this purpose. After that, we generate the contextual counterfactual for these syntactically similar entities to meet the condition of invariant entities, i.e., instead of changing entities, we intervene on their contexts. We term this proposed method **SynCo** as the substitution of causal features is based on the syntactic graph SynDG.

We take two samples for illustration: an original sample X_{ori} "They drank wine produced by wineries." with "Product-Producer" relation, and a candidate sample X_{can} "They bought the grapes from farms" with "Entity-Origin". Their SynDGs and the intervention procedure are shown in Figure 2.



Sample	Candidate Term	BC	CC	DC	AvgC
Original sample	wine (e)	0.6	1.11	2	1.24
	wineries (e)	0.0	0.67	1	0.56
	drank(v)	0.6	0.91	2	1.17
	produced (v)	0.6	1.11	2	1.24
Candidate sample	grapes (e)	0.4	1.11	2	1.17
	farms (e)	0.0	0.67	1	0.56
	bought (v)	0.8	1.43	3	1.74

BC : Betweenness Centrality ; CC : Closeness Centrality ; DC : Degree Centrality.

Figure 2: An illustration of SynCo. The left is SynDG of two samples and the topological distance between entities. The right is the result for centrality calculation. TDT = 0.2.

The main procedure for SynCo is summarized below.

1. We first calculate the average score for three centrality metrics of the entity and denote it as avgC(entity), e.g., avgC(wine_o) and avgC(grapes_c) is 1.24 and 1.17.

2. We then calculate the topological distance (TD) of two entities by minusing their avgC values. If TD is smaller than a predefined topological distance threshold (TDT), they form a candidate entity pair for substitute structures. For example, given TD(avgC(wine_o), avgC(grapes_c)) = 0.07 < TDT and TD(avgC(wine_o), avgC(farms_c)) = 0.68 > TDT, the entity 'grapes' in X_{can} and the entity 'wine' in X_{ori} form a candidate entity pair.

3. We now generate the candidate syntactic counterfactual. We calculate the cosine similarity of the syntactic features (POS and dependency relation embedding) for entities in the candidate entity pair. We denote it as FS. If it is greater than a predefined feature similarity threshold (FST), we substitute the first-order neighbors around the entity in X_{ori} with those of the candidate entity in X_{can}. These neighbors should have the same type of POS tags to ensure the correctness of the syntax. Moreover, if there are several words around entities and they all have the same type, we will replace the one with the smallest TD. For example, given FS(wine_o, grapes_c) = 0.877 > FST, we can substitute the verb-type neighbors 'drank_o' and 'produced_o' of 'wine' with the same type neighbor 'bought_c' of 'grapes'. Moreover, since TD(avgC(drank_o), avgC(bought_c)) = 0.57 and TD(avgC(produced_o), avgC(bought_c)) = 0.50, we replace 'produced_o' with 'bought_c' and retain 'drank_o' unchanged.

[3]We randomly select three substitutes for the original sample. If none of generated candidates meets the requirements, they will all be discarded. If multiple candidates are qualified as counterfactuals, one of them is chosen at random.

4. After substituting the contexts of the candidate entity pair and adding the label of X_{can}, a candidate counterfactual is produced. In order to ensure that the sample we generate is a real counterfactual, we put it into the trained backbone model and re-predict its label. If the label is indeed changed, we treat it as the counterfactual of X_{ori}. For example, "They drank wine bought from wineries." with "Entity-Origin" relation is a counterfactual sample after SynCo.

4.2 Generating Semantic Counterfactuals

This section presents our semantically intervened counterfactual generator **SemCo**. We prefer to replace the contexts between entities in X_{ori} with words of the most similar semantics and get a different label. To this end, we exploit the shortest path in SemDG for identifying semantically similar contexts between entities.

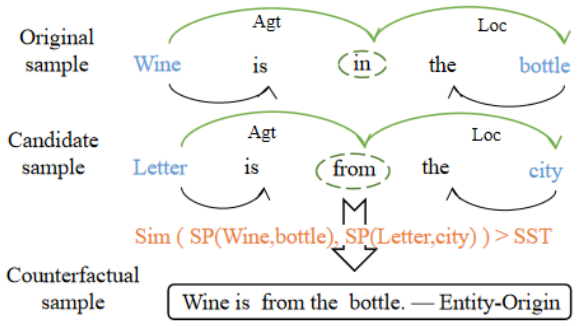


Figure 3: An illustration of SemCo. SP(e1,e2): the shortest semantic path between two entities. SST = 0.6.

As shown in Figure 3, we first obtain the SemDG of an original sample "Wine is in the bottle" with "Content-Container" label and that of a candidate sample "Letter is from the city" with "Entity-Origin" label. We then extract the SDP between entities in X_{ori} and X_{can}. We propose to calculate the cosine similarity of averaged word embeddings between two paths. If the similarity score is larger than the semantic similarity threshold (SST), we replace the semantic path in X_{ori} with that in X_{can}. Similar to SynCo, we put the generated semantic counterfactual into the trained backbone model and re-predict its label. We consider the new sample as a counterfactual of X_{ori} if the label is changed.

Finally, the generated semantic counterfactuals, as well as the syntactic counterfactuals, are used to augment the original data to train a robust classifier.

5 DATASETS AND EVALUATION PROTOCOL

5.1 Datasets

We adopt three in-domain datasets on two RE tasks, including **PubMed** [17] for cross-sentence n-ary RE, and **TACRED** [38] and **SemEval** [6] for sentence-level RE. We also employ one out-of-domain dataset **ACE2005** [25] to verify the model robustness with CAD. More details about datasets are given in Appendix.

5.2 Evaluation Protocol

We evaluate our model with a dedicated protocol.

1. Backbones. We first adopt three typical RE methods as the backbones, which are used as the encoder for the input and for training the base classifier for the prediction of candidate counterfactuals. Moreover, the backbones are used for training the final classifier on the original data and CAD. The backbones include PA-LSTM [38], AGGCN [5], and R-BERT [29]. They are chosen as the representative of sequence based, graph-based, and PLM based methods, respectively. These backbones will be combined with all counterfactual generation methods. We also show improvements over joint entity and relation extraction backbones like PURE [39] and UniRE [27] under a different evaluation protocol in Appendix.

2. Baselines. We further compare our model with automatical counterfactual generation baselines, including general text generator GYC [13], CLOSS [4], CF-GAN [21], and the specific generator originally designed for other tasks like COSY [36], RM-CT [34], and Gcc [11].

For general methods GYC [13], CLOSS [4], and CF-GAN [21], we directly generate counterfactuals on RE datasets using their generator, and we avoid masking entities for a fair comparison with our model. Among the specific methods, COSY [36] is for the cross-lingual understanding task, and we follow it to build a counterfactual dependency graph by maintaining the graph structure, and replacing each type of relation with a randomized one and randomly selecting a POS tag. RM-CT [34] is proposed for sentiment classification tasks. We employ its self-supervised context decomposition and select the human-like counterfactuals using MoverScore. Gcc [11] generates different relation instances from specifically designed relation triplets and we use WordNet for its relation renaming strategy.

For models with static word embedding, we initialize word vectors with 300-dimension GloVe embeddings provided by Pennington et al. [18]. For models with contextualized word embedding, we adopt BERT_{base} as the PLM. The embeddings for POS and dependency relation are initialized randomly and their dimension is set to 30. We adopt the Stanford CoreNLP [15] as the syntactic parser and use semantic dependency parsing [3] to generate semantic dependency graph. We use SGD as the optimizer with a 0.9 decay rate. The L2-regularization coefficient λ is 0.002. We re-produce the results for baselines (RM-CT, CLOSS, CF-GAN, and Gcc) with the source code provided by the authors. For methods without released code like COSY and GYC, we implement them by ourselves using the optimal hyper-parameter settings reported in their papers. Below is the detailed process.

(1) Data sampling. CLOSS generates counterfactuals by sampling about 1000 training data according to the original paper. RM-CT tries to generate counterfactuals using a large number of samples, e.g., all the data or 1000 samples for each class, but the results are not very good due to the introduction of low-quality counterfactuals. We test several cases and choose the best setting (50 examples of each relation) from the training set to generate counterfactual data. CF-GAN and GYC generate counterfactuals for each data according to the settings in their original papers, the counterfactual for each instance is selected by generating the top-5 counterfactuals and selecting the one closest to target sample measured by the cosine similarity. Gcc generates counterfactuals using a large number of samples, e.g., all the data, but the results are not very good due to the introduction of low-quality counterfactuals. We test several cases which use 500, 1000, 2000 samples in total or 50 samples for each class like RM-CT from the training data, and finally choose

Table 1: The number of augmented samples and the ratio of label-flipped samples after re-prediction.

Data	PubMed_T		PubMed_B		SemEval		TACRED	
	Num.	Ratio	Num.	Ratio	Num.	Ratio	Num.	Ratio
PA-LSTM								
COSY	5313	0	4613	0	8000	0	68124	0
RM-CT	264	58%	216	46%	145	30%	41	5%
GYC	5313	100%	4613	100%	8000	100%	68124	100%
CLOSS	1000	100%	1000	100%	1000	100%	1000	100%
CF-GAN	5313	100%	4613	100%	8000	100%	68124	100%
Gcc	2000	100%	2000	100%	2000	100%	2000	100%
CoCo	1452	42%	1320	39%	2189	51%	12457	29%
AGGCN								
COSY	5313	0	4613	0	8000	0	68124	0
RM-CT	15	5%	31	8%	21	5%	648	35%
GYC	5313	100%	4613	100%	8000	100%	68124	100%
CLOSS	1000	100%	1000	100%	1000	100%	1000	100%
CF-GAN	5313	100%	4613	100%	8000	100%	68124	100%
Gcc	2000	100%	2000	100%	2000	100%	2000	100%
CoCo	1547	45%	1389	41%	2418	56%	15645	31%
R-BERT								
COSY	5313	0	4613	0	8000	0	68124	0
RM-CT	641	45%	628	46%	154	32%	962	41%
GYC	5313	100%	4613	100%	8000	100%	68124	100%
CLOSS	1000	100%	1000	100%	1000	100%	1000	100%
CF-GAN	5313	100%	4613	100%	8000	100%	68124	100%
Gcc	2000	100%	2000	100%	2000	100%	2000	100%
CoCo	1475	49%	1421	44%	2426	61%	18269	37%

the best setting, i.e., 1000 samples which are used to generate 1000 positive and 1000 negative samples for its contrastive learning.

(2) Selecting the target label. During the process of generating counterfactuals, some methods need to determine the flipped labels in advance. In RE, when the original backbone is used for classification, two labels, which have the second and third highest probability values predicted by the classifier of the backbone, are selected as the target labels to be appended to the counterfactual sample. Note the label with the highest probability value is used as the original label.

(3) Generation. Generating counterfactuals according to the settings in their original papers.

6 RESULTS AND DISCUSSIONS

We first present the statistic analysis to get an overview of proposed method and other baselines. We then compare our method with the backbone and the state-of-art automatical counterfactual generation methods on three in-domain datasets. We also provide an ablation study to focus on the contribution of single component in our model. Furthermore, we present a case study of counterfactuals for a detailed comparison of the counterfactuals generated by baselines and our model. Notably, we evaluate our model on the out-of-domain data for the generalization test. We finally compare the impact of the counterfactuals and that of label invariant samples.

Due to the space limitation, we omit the parameter analysis on topological distance threshold (TDT), feature similarity threshold (FST), and semantics similarity threshold (SST) for our model, and present these results in Appendix.

Table 2: Main results in terms of accuracy on PubMed. "T" and "B" denote the ternary and binary entity interactions, and "Single" and "Cross" mean the accuracy calculated within single sentences or on all sentences. The best results are in bold, and the second best ones are underlined. $\hat{}$ and $*$ mark denote statistically significant improvements over the backbone results with $p < .05$ and $p < .01$, and † and ‡ mark denote statistically significant improvements over the corresponding second best results with $p < .05$ and $p < .01$, respectively. $_{bb}$ denotes 'backbone'.

Model	Binary-class				Multi-class	
	T		B		T	B
	Single	Cross	Single	Cross	Cross	Cross
PA-LSTM$_{bb}$	84.9	85.8	85.6	85.0	78.1	77.0
+ COSY	85.2	86.1	85.8	85.4	78.4	77.3
+ RM-CT	85.0	84.9↓	84.9↓	85.0	77.7↓	76.9↓
+ GYC	85.1	85.9	85.4↓	85.3	78.2	77.2
+ CLOSS	84.9↓	84.6↓	85.7	85.2	77.4↓	76.9↓
+ CF-GAN	84.6↓	85.7↓	85.1↓	85.2	78.0↓	77.1
+ Gcc	84.9	85.9	85.7	85.2	78.2	77.1
+ **CoCo**	**87.0***‡	**86.9**$^{\hat{}}$	**87.5**$^{\hat{}†}$	**86.8***‡	**80.5***‡	**80.0***‡
AGGCN$_{bb}$	87.1	87.0	85.2	85.6	79.7	77.4
+ COSY	87.6	87.8	86.3	86.2	80.7	78.4
+ RM-CT	87.1	87.0	85.3	85.5↓	79.7	77.0↓
+ GYC	87.2	87.1	85.4	85.6	79.8	77.6
+ CLOSS	76.4↓	86.3↓	85.0↓	84.3↓	79.2↓	76.4↓
+ CF-GAN	87.1	87.1	85.1	85.4	80.1	77.2
+ Gcc	87.2	87.1	85.3	85.6	79.8	77.5
+ **CoCo**	**89.0***‡	**89.1***‡	**88.0***‡	**87.7***†	**84.1***‡	**81.1***‡
R-BERT$_{bb}$	88.6	88.7	88.1	87.9	85.1	84.2
+ COSY	88.6	88.8	88.3	87.9	85.3	84.5
+ RM-CT	88.5	88.6↓	87.9↓	87.6↓	85.0↓	84.2
+ GYC	88.7	88.9	88.0↓	87.8↓	85.2	84.3
+ CLOSS	87.2↓	87.5↓	87.1↓	86.7↓	83.8↓	83.4↓
+ CF-GAN	88.1↓	88.2↓	87.4↓	87.3↓	84.8↓	84.1↓
+ Gcc	88.6	88.8	88.2	87.8↓	85.1	84.3
+ **CoCo**	**89.1**$^{\hat{}}$	**89.3**$^{\hat{}}$	**88.7**$^{\hat{}}$	**88.4**$^{\hat{}}$	**86.2**$^{\hat{}†}$	**85.8**$^{\hat{}†}$

6.1 Statistic Analysis

We present the number of the augmented samples and the ratio of label-flipped samples after re-prediction for both proposed method and baselines in Table 1.

(1) GYC, CLOSS, CF-GAN, and Gcc can generate the same number of samples as those in training data (CLOSS and GCC has a threshold of 1000 and 2000, respectively), and the label changed ratio is 100% since they pre-determine the label before generation. However, such a policy may introduce low-quality counterfactuals, as we will see in main comparison results.

(2) COSY also generates the same number of samples as those in training data, and its label changed ratio is 0%. It is actually a traditional DA method since it randomly replaces syntactic features which cannot flip the label.

(3) RM-CT generates the smallest number of counterfactuals as it tries to change the label of samples by removing words in sentences. This operation may destroy the syntax structure and cannot change the label in many cases.

(4) The generation process of baselines is often uncontrollable and it is easy for them to produce sentences with lexical or syntactic errors or non-counterfactuals. In contrast, we carefully design the

Table 3: Main results on TACRED and SemEval. Micro-avg. precision (P), recall (R), and F1 on TACRED. Macro-avg. F1 on SemEval. The marks are as same as those in Table 2.

Model	TACRED			SemEval
	P	R	Micro-F1	Macro-F1
PA-LSTM$_{bb}$	65.7	64.5	65.1	82.7
+ COSY	65.8	64.6	65.2	83.1
+ RM-CT	**66.9**	63.3↓	65.0↓	80.1
+ GYC	65.2↓	64.1↓	64.6↓	82.9
+ CLOSS	64.2↓	63.9↓	64.0↓	81.3↓
+ CF-GAN	64.9↓	63.7↓	64.3↓	82.4↓
+ Gcc	65.8	64.5	65.2	82.8
+ **CoCo**	66.3	**66.1**	**66.2***†	**84.2***‡
+ AGGCN$_{bb}$	71.9	64.0	67.7	85.7
+ COSY	71.8↓	64.2	67.8	85.9
+ RM-CT	71.0↓	63.9↓	67.6↓	84.8↓
+ GYC	71.3↓	63.9↓	67.4↓	85.6↓
+ CLOSS	70.2↓	63.3↓	66.6↓	84.8↓
+ CF-GAN	71.0↓	63.3↓	66.9↓	85.4↓
+ Gcc	71.9	64.3	67.9	85.9
+ **CoCo**	**72.4**	**64.8**	**68.4**$^{\hat{}\dagger}$	**86.6**$^{\hat{}\dagger}$
R-BERT$_{bb}$	69.7	70.1	69.9	88.6
+ COSY	69.6↓	70.1	69.8↓	88.5↓
+ RM-CT	69.2↓	69.8↓	69.5↓	87.2↓
+ GYC	69.5↓	70.5	70.0	88.6
+ CLOSS	68.0↓	67.9↓	67.9↓	87.5↓
+ CF-GAN	68.9↓	69.3↓	69.1↓	87.5↓
+ Gcc	69.7	70.2	70.0	88.7
+ **CoCo**	**70.2**	**70.5**	**70.4**	**89.0**

strategy for counterfactual generation. As a result, CoCo generates about 30%∼60% label-changed high quality counterfactuals for each dataset.

6.2 Main Results

The main results for cross-sentence *n*-ary and sentence-level tasks are shown in Table 2 and Table 3, respectively. From these results, we make the following observations.

(1) Our CoCo model significantly improves the performance of all the backbones across three datasets. Specifically, CoCo outperforms two backbone approaches PA-LSTM and AGGCN by around 2-3 absolute percentage points in terms of accuracy on PubMed. It also improves F1 scores by 1-2 absolute percentage points on TACRED and SemEval. More importantly, our model boosts the performance of R-BERT on all three datasets especially on PubMed, which is impressive because it is a very strong backbone, and also because most baselines damage the performance of R-BERT.

(2) Our model achieves the state-of-the-art performance in terms of a counterfactual generator for RE. For example, CoCo outperforms the syntax-based model COSY and the best semantics-based model GYC by 1-3 absolute percentage points. COSY randomly replaces syntactic features which cannot flip the label and makes little sense for RE. CF-GAN and GYC get a slight increase on PubMed with PA-LSTM and AGGCN, but their other results are unsatisfactory. The results of RM-CT on three datasets fluctuate. On PubMed, some of them keep the same as the original ones and some of them decline, while on SemEval and TACRED they all decline. The poor results of these baseline counterfactual generators can be due to the lack of grammar constraint and the lack of entity-centric viewpoint. The performance of Gcc is better on SemEval and TACRED than that on PubMed. The reason might be that it requires external knowledge like WordNet to build prototypical relation instance. For biomedical relations in PubMed, it is hard to find accurate alias relation names to generate positive examples and thus Gcc gets worse results.

Table 4: Ablation results. (_B) and (_T) denote the binary and ternary relation on PubMed.

Model	PubMed_B Acc.	PubMed_T Acc.	TACRED Micro-F1	SemEval Macro-F1
PA-LSTM$_{bb}$	77.0	78.1	65.1	82.7
+ SynCo	79.4	79.6	65.6	83.9
+ SemCo	**79.7**	**79.8**	**65.7**	**84.0**
+ Syn-TED	77.6	78.4	65.2	83.0
+ Sem-BA	77.4	78.3	65.3	83.2
AGGCN$_{bb}$	77.4	79.7	67.7	85.7
+ SynCo	80.3	82.8	67.9	86.3
+ SemCo	**80.7**	**83.1**	**68.2**	**86.5**
+ Syn-TED	77.6	79.9	67.7	85.8
+ Sem-BA	77.8	79.8	67.8	86.1
R-BERT$_{bb}$	84.2	85.1	69.9	88.6
+ SynCo	85.1	85.6	70.1	**88.7**
+ SemCo	**85.6**	**85.8**	**70.3**	**88.7**
+ Syn-TED	84.3	85.1	70.0	88.5
+ Sem-BA	84.6	85.2	70.0	88.3

6.3 Ablation Study

Our model has two unique characteristics. Firstly, it utilizes both syntactic and sematic information. Secondly, it takes advantage of the graph topological property. We hence design two types of ablation study. One is performing one separate component only, i.e., SynCo or SemCo. The other is replacing our graph topology based intervention methods with other syntactic/semantic ones, i.e., tree edit distance (TED) [37] which measures the syntactic closeness between the candidate and the original text and BERT-attack (BA) [10] which generates substitutes for the vulnerable words in a semantic-preserving way. We present the results on three datasets in Table 4.

We find that a single SynCo or SemCo is already good enough to enhance the performance of the backbone. In addition, we observe that our single SynCo/SemCo outperforms the baselines in Table 2 and Table 3. Moreover, SynCo and SemCo have almost the same effects on the model. For example, the backbone PA-LSTM has an accuracy score 77.0 on PubMed_B. After SynCo/SemCo, its accuracy rises up to 79.4/79.7, showing a similar 2.4/2.7 absolute increase. These results demonstrate that both SynCo and SemCo contribute to our model, and their combination CoCo is more powerful.

The replacement of SynCo with TED and SemCo with BA hurts the performance. For example, on SemEval with AGGCN as the backbone, TED results in a 0.5 F1 decrease (SynCo 86.3 *vs.* Syn-TED 85.8). On TACRED with R-BERT as the backbone, BA brings about a 0.3 (SemCo 70.3 *vs.* Syn-TED 70.0) absolute decrease. The reason might be that TED only considers the causal words from the whole syntactic tree while our SynCo exploits the topological structure and syntactic feature of the words. Meanwhile, BA is unable to capture causal associations when generating adversarial samples

Table 5: Case study. Entities are in bold and the words in red represent newly generated words.

Case1	The sisters are teenage **refugees** from a violent **home**. — Entity-Origin
RM-CT	The sisters are teenage **refugees** ~~from~~ a violent **home**. — Other
CF-GAN	The sisters are teenage **refugees** ~~from~~ sleep a violent **home**. — Member-Collection
Gcc	Pos: **refugees** are from a violent **home**. — Entity-from Neg: **refugees** are eager to return **home**. — Entity-Destination
CoCo	The sisters are teenage **refugees** ~~from~~ lived in a violent **home**. — Member-Collection
Case2	the **wounds** caused by the **scour ging** and the thorns are almost invisible. — Cause-Effect
CLOSS	the **wounds** ~~caused~~ community by the **scour ging** and the thorns are almost invisible. — Product-Producer
GYC	the **wounds** ~~caused by~~ develop the **scour ging** and the thorns are ~~almost~~ very invisible. — Product-Producer
Gcc	Pos: He got the **wounds** because of the **scour ging**. — Reason Neg: **scour ging** produced the **wounds** badly. — Product-Producer
CoCo	the **wounds** ~~caused~~ produced by **scour** ging and the thorns are almost invisible. — Product-Producer

since it only replaces the words with the similar ones generated by BERT.

6.4 Case Study

To have a close look, we select two samples from SemEval for case study and present results by different models in Table 5. RM-CT deletes the important preposition 'from' to change the label and results in an incomplete sample. CF-GAN employs an adversarial attack method. Its generated 'sleep' is ungrammatical and the sample is low-quality as it is inconsistent with the flipped label. GYC and CLOSS generate the counterfactual with a specified target label and the substitution tends to be common words in the specified class like 'community', which makes the sentence unreadable. Moreover, the word 'almost' identified by GYC is out of the scope of two entities and is not a causal feature.

The results by the above baselines are very uncontrollable, and it is easy for them to generate sentences with semantic or syntactic errors. Gcc is good and produces reasonable samples in most cases except the one '...produced the wounds badly'. However, Gcc cannot explicitly identify the causal features during its generation process and thus the model lacks interpretability. In contrast, our model not only recognizes the correct causal features, but also produces human-like counterfactuals, which forces the classifier to better distinguish causation and confounding.

6.5 Robustness in the Generalization Test

To evaluate the model robustness, we perform the generalization test on ACE2005 dataset by using the training and test data from different domains. Following the settings for fine-grained RE tasks [35], we use the union of the news domains (nw and bn) for training, and hold out half of the bc domain as development data, and finally evaluate on the remainder of bc, cts, and wl domains. We choose three best baselines GYC, Gcc, and COSY for comparison.

As can be seen in Table 6, our proposed CoCo model enhances the performance of all backbones on the combined CAD and the original (Ori.) data, i.e., PA-LSTM (+4.1%, +3.7%, +2.3%), AGGCN (+1.9%, +2.2%, +1.8%), and R-BERT (+0.5%, +1.0%, +0.7%) on three different target domains. It is worth noting that our improvements over three backbones (including R-BERT) are statistically significant on all generalization tests. Another interesting finding is that COSY is the worst in almost all cases in this generalization test while it is usually better than other two baselines on in-domain data. Note our CoCo and GYC and GCC all flip the labels by changing the causal features. This is a convincing evidence of the link between causal features and the model robustness.

Table 6: Results for the generalization test on ACE2005.

Different Training Data	PA-LSTM	AGGCN	R-BERT
Micro-avg. F1 on **bc** domain.			
Ori.	48.5	62.5	68.5
Ori. & CAD (GYC)	48.7	63.1	68.6
Ori. & CAD (Gcc)	48.7	63.2	68.9
Ori. & CAD (COSY)	48.6	62.4 ↓	68.6
Ori. & CAD (CoCo)	**52.6***‡	**64.4***†	**69.0**$^{\hat{}}$
Micro-avg. F1 on **cts** domain.			
Ori.	42.5	63.1	69.4
Ori. & CAD (GYC)	42.6	63.8	69.8
Ori. & CAD (Gcc)	42.6	63.3	69.7
Ori. & CAD (COSY)	42.3 ↓	63.2	69.7
Ori. & CAD (CoCo)	**46.2***‡	**65.3***†	**70.4**$^{\hat{}}$
Micro-avg. F1 on **wl** domain.			
Ori.	38.8	53.4	59.5
Ori. & CAD (GYC)	39.1	53.6	59.7
Ori. & CAD (Gcc)	39.0	53.7	59.8
Ori. & CAD (COSY)	38.1 ↓	53.7	59.2 ↓
Ori. & CAD (CoCo)	**41.1***‡	**55.2***‡	**60.2**$^{\hat{}}$

6.6 Comparison between Counterfactuals and Label Invariant Samples

Several previous studies [30, 36] do not differentiate label invariant and label flipped samples and treat all these samples as counterfactuals. To clarify the problem, we develop a CoCo^{-} method which has the same procedure as CoCo but it selects the label invariant samples at the final step. By doing this, we wish to investigate the impact of label invariant samples and that of label flipped samples (counterfactuals) under the same generation condition.

We first present the number of the augmented samples and the ratio of label invariant/flipped samples after re-prediction in Table 7. Note we generate three pseudo samples for each candidate sample in the training data, and re-predict their labels using the backbone method. Each candidate sample may generate 0 to 3 label invariant/flipped pseudo sample(s), and we will choose at most one label invariant or label flipped pseudo sample to augment the training data, respectively. We divide the total number of label invariant/flipped pseudo samples into all candidate samples to calculate the ratio.

Table 7: The number of augmented samples and the ratio of label invariant/flipped samples after re-prediction.

Data	PubMed_T		PubMed_B		SemEval		TACRED	
	Num.	Ratio	Num.	Ratio	Num.	Ratio	Num.	Ratio
PA-LSTM								
CoCo$^-$	3861	58%	3293	61%	5811	49%	55667	71%
CoCo	1452	42%	1320	39%	2189	51%	12457	29%
AGGCN								
CoCo$^-$	3766	55%	3242	59%	5582	44%	52479	69%
CoCo	1547	45%	1389	41%	2418	56%	15645	31%
R-BERT								
CoCo$^-$	3838	51%	3192	56%	5574	39%	49855	63%
CoCo	1475	49%	1421	44%	2426	61%	18269	37%

Table 8: Comparison results in terms of accuracy between the label invariant samples and the counterfactuals on the in-domain PubMed dataset. The markers are as same as those in Table 2.

Model	Binary-class				Multi-class	
	T		B		T	B
	Single	Cross	Single	Cross	Cross	Cross
PA-LSTM$_{bb}$	84.9	85.8	85.6	85.0	78.1	77.0
+ CoCo$^-$	85.4	86.2	85.9	85.6	78.4	77.2
+ **CoCo**	**87.0***‡	**86.9**$^{\hat{}}$	**87.5**$^{\hat{}\dagger}$	**86.8***†	**80.5***‡	**80.0***‡
AGGCN$_{bb}$	87.1	87.0	85.2	85.6	79.7	77.4
+ CoCo$^-$	87.7	87.8	86.4	86.4	80.7	78.6
+ **CoCo**	**89.0***‡	**89.1***‡	**88.0***‡	**87.7***†	**84.1***‡	**81.1***‡
R-BERT$_{bb}$	88.6	88.7	88.1	87.9	85.1	84.2
+ CoCo$^-$	88.7	88.8	88.4	88.0	85.2	84.5
+ **CoCo**	**89.1**$^{\hat{}}$	**89.3**$^{\hat{}}$	**88.7**$^{\hat{}}$	**88.4**$^{\hat{}}$	**86.2**$^{\hat{}\dagger}$	**85.8**$^{\hat{}\dagger}$

Table 9: Comparison results between the label invariant samples and the counterfactuals on the in-domain TACRED and SemEval datasets.

Model	TACRED			SemEval
	P	R	Micro-F1	Macro-F1
PA-LSTM$_{bb}$	65.7	64.5	65.1	82.7
+ CoCo$^-$	65.8	64.6	65.3	83.0
+ **CoCo**	**66.3**	**66.1**	**66.2***†	**84.2***‡
+ AGGCN$_{bb}$	71.9	64.0	67.7	85.7
+ CoCo$^-$	71.9	64.2	67.9	86.0
+ **CoCo**	**72.4**	**64.8**	**68.4**$^{\hat{}\dagger}$	**86.6**$^{\hat{}\dagger}$
R-BERT$_{bb}$	69.7	70.1	69.9	88.6
+ CoCo$^-$	69.7	70.1	69.9	88.7
+ **CoCo**	**70.2**	**70.5**	**70.4**	**89.0**

It is clear from Table 7 that the absolute number of label invariant samples is larger than that of label flipped ones. This is reasonable since in a sentence, most features (words or tokens) are irrelevant contextual words (spurious features), and only a small number of them are causal features. It is a bit strange that the label flipped counterfactuals has a large ratio yet a small number on SemEval. This indicates that the counterfactuals are concentrated on some sentences in this dataset, which may produce 2 or 3 counterfactuals. However, we only choose one of them to augment the original data.

We then compare the experimental results by adding the label invariant samples and the counterfactuals on three in-domain datasets including PubMed, TACRED, and SemEval in Table 8 and Table 9, respectively. We can see that, on all three in-domain datasets, the counterfactuals have more positive impacts than the label invariant samples though the size of augmented data for counterfactuals is much smaller than that of label invariant samples.

Table 10: Comparison results between the label invariant samples and the counterfactuals on the out-of-domain ACE2005 dataset.

Different Training Data	PA-LSTM	AGGCN	R-BERT
Micro-avg. F1 on **bc** domain.			
Ori.	48.5	62.5	68.5
Ori. & CAD (CoCo$^-$)	48.6	62.6	68.6
Ori. & CAD (CoCo)	**52.6***‡	**64.4***‡	**69.0**$^{\hat{}}$
Micro-avg. F1 on **cts** domain.			
Ori.	42.5	63.1	69.4
Ori. & CAD (CoCo$^-$)	42.4$\downarrow$	63.3	69.6
Ori. & CAD (CoCo)	**46.2***‡	**65.3***‡	**70.4**$^{\hat{}\dagger}$
Micro-avg. F1 on **wl** domain.			
Ori.	38.8	53.4	59.5
Ori. & CAD (CoCo$^-$)	38.9	53.6	59.3$\downarrow$
Ori. & CAD (CoCo)	**41.1***‡	**55.2***‡	**60.2**$^{\hat{}\dagger}$

We finally compare their ability of enhancing the model robustness on the out-of-domain dataset in Table 10. It is clear that CoCo$^-$ has a poor generalization ability. Similar to COSY, CoCo$^-$ sometimes has a negative impact on the backbone model with augmented label invariant samples. In contrast, CoCo always improves the backbone with the augmented label flipped counterfactuals. All these results prove that counterfactuals play an important role to enhance the model robustness and the change of labels is critical to the recognition of causal features.

7 CONCLUSION

In this paper, we introduce the problem of automatic counterfactual generation into the RE task. We aim to produce the most human-like, i.e., grammatically correct and semantically readable, counterfactuals, while keeping the entities unchanged. To this end, we design an entity-centric framework which employs semantic and syntactic dependency graphs and exploits two topological properties in these two graphs to first identify and then intervene on contextual causal features for entities. Extensive experimental results prove that our model significantly outperforms the backbones, and it is also more effective in alleviating spurious associations and improving the model robustness than the state-of-the-art baselines. Since our method generates counterfactuals by replacing causal features with those from other classes in the training data, the number of generated counterfactuals is relatively small.

In the future, we will combine our model with PLM based methods to obtain more counterfactuals and increase the diversity. Moreover, the ratio of errors produced by the backbone has impacts on the quality of CAD, which is the limitation of our model. We plan to solve this problem by introducing external knowledge.

ACKNOWLEDGMENTS

The work was supported by a grant from the National Natural Science Foundation of China (NSFC) (Grant No. 62276193).

REFERENCES

[1] Hao Chen, Rui Xia, and Jianfei Yu. 2021. Reinforced Counterfactual Data Augmentation for Dual Sentiment Classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 269–278. https://doi.org/10.18653/v1/2021.emnlp-main.24

[2] Xiang Chen, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. KnowPrompt: Knowledge-aware Prompt-tuning with Synergistic Optimization for Relation Extraction. In *Proceedings of the ACM Web Conference 2022*. ACM, Virtual Event, Lyon, France, 2778–2788. https://dl.acm.org/doi/10.1145/3485447.3511998

[3] Timothy Dozat and Christopher D. Manning. 2018. Simpler but More Accurate Semantic Dependency Parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 484–490. https://doi.org/10.18653/v1/P18-2077

[4] Xiaoli Z. Fern and Quintin Pope. 2021. Text Counterfactuals via Latent Optimization and Shapley-Guided Search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 5578–5593. https://doi.org/10.18653/v1/2021.emnlp-main.452

[5] Zhijiang Guo, Yan Zhang, and Wei Lu. 2019. Attention Guided Graph Convolutional Networks for Relation Extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 241–251. https://doi.org/10.18653/v1/p19-1024

[6] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 Task 8: Multi-Way Classification of Semantic Relations between Pairs of Nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*. 33–38. https://aclanthology.org/S10-1006/

[7] Jason Jo and Yoshua Bengio. 2017. Measuring the tendency of CNNs to Learn Surface Statistical Regularities. *CoRR* abs/1711.11561 (2017). http://arxiv.org/abs/1711.11561

[8] Pearl Judea. 2000. *Causality: models, reasoning, and inference*. Vol. 41. Cambridge University Press, 189–190. https://doi.org/10.1016/S0925-2312(01)00330-7

[9] Divyansh Kaushik, Eduard H. Hovy, and Zachary Chase Lipton. 2020. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Sklgs0NFvr

[10] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 6193–6202. https://doi.org/10.18653/v1/2020.emnlp-main.500

[11] Fangchao Liu, Lingyong Yan, Hongyu Lin, Xianpei Han, and Le Sun. 2021. Element Intervention for Open 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 1: Long Papers)*. 4683–4693.

[12] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3219–3232. https://doi.org/10.18653/v1/D18-1360

[13] Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate Your Counterfactuals: Towards Controlled Counterfactual Generation for Text. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 13516–13524. https://ojs.aaai.org/index.php/AAAI/article/view/17594

[14] Angrosh Mandya, Danushka Bollegala, and Frans Coenen. 2020. Graph Convolution over Multiple Dependency Sub-graphs for Relation Extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*. 6424–6435. https://doi.org/10.18653/v1/2020.coling-main.565

[15] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 55–60. https://doi.org/10.3115/v1/p14-5010

[16] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of the ACM Web Conference 2020*. ACM, Taipei, Taiwan. https://dl.acm.org/doi/abs/10.1145/3366423.3380087

[17] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-Sentence N-ary Relation Extraction with Graph LSTMs. *Trans. Assoc. Comput. Linguistics* 5 (2017), 101–115. https://doi.org/10.1162/tacl_a_00049

[18] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543. https://doi.org/10.3115/v1/d14-1162

[19] Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: Improving Entity and Relation Understanding for Pre-trained Language Models via Contrastive Learning. 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)*. 3350–3363. https://doi.org/10.18653/v1/2021.acl-long.260

[20] Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling Relations and Their Mentions without Labeled Text. In *Machine Learning and Knowledge Discovery in Databases*, José Luis Balcázar, Francesco Bonchi, Aristides Gionis, and Michèle Sebag (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 148–163.

[21] Marcel Robeer, Floris Bex, and Ad Feelders. 2021. Generating Realistic Natural Language Counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 3611–3625. https://doi.org/10.18653/v1/2021.findings-emnlp.306

[22] Dan Roth and Wen-tau Yih. 2004. A Linear Programming Formulation for Global Inference in Natural Language Tasks. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004*. Association for Computational Linguistics, Boston, Massachusetts, USA, 1–8. https://aclanthology.org/W04-2401

[23] Megha Srivastava, Tatsunori B. Hashimoto, and Percy Liang. 2020. Robustness to Spurious Correlations via Human Annotations. In *International Conference on Machine Learning*, Vol. 119. 9109–9119. http://proceedings.mlr.press/v119/srivastava20a.html

[24] Mycal Tucker, Peng Qian, and Roger Levy. 2021. What if This Modified That? Syntactic Interventions with Counterfactual Embeddings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 862–875. https://doi.org/10.18653/v1/2021.findings-acl.76

[25] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. In *LDCP*.

[26] 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 *Proc. of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.

[27] Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. 2021. UniRE: A Unified Label Space for Entity 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 1: Long Papers)*. Association for Computational Linguistics, Online, 220–231. https://doi.org/10.18653/v1/2021.acl-long.19

[28] Zhao Wang and Aron Culotta. 2021. Robustness to Spurious Correlations in Text Classification via Automatically Generated Counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 14024–14031. https://ojs.aaai.org/index.php/AAAI/article/view/17651

[29] Shanchan Wu and Yifan He. 2019. Enriching Pre-trained Language Model with Entity Information for Relation Classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 2361–2364. https://doi.org/10.1145/3357384.3358119

[30] Tongshuang Wu, Marco Túlio Ribeiro, Jeffrey Heer, and Daniel S. Weld. 2021. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. 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)*. 6707–6723. https://doi.org/10.18653/v1/2021.acl-long.523

[31] Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Paths. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 1785–1794. https://doi.org/10.18653/v1/d15-1206

[32] 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)*. 6442–6454. https://doi.org/10.18653/v1/2020.emnlp-main.523

[33] Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. A Partition Filter Network for Joint Entity and Relation Extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 185–197. https://aclanthology.org/2021.emnlp-main.17

[34] Linyi Yang, Jiazheng Li, Padraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. Exploring the Efficacy of Automatically Generated Counterfactuals for Sentiment Analysis. 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)*. 306–316. https://doi.org/10.18653/v1/2021.acl-long.26

[35] Mo Yu, Matthew R. Gormley, and Mark Dredze. 2015. Combining Word Embeddings and Feature Embeddings for Fine-grained Relation Extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1374–1379. https://doi.org/10.3115/v1/n15-1155

[36] Sicheng Yu, Hao Zhang, Yulei Niu, Qianru Sun, and Jing Jiang. 2021. COSY: COunterfactual SYntax for Cross-Lingual Understanding. 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)*. 577–589. https://doi.org/10.18653/v1/2021.acl-long.48

[37] Kaizhong Zhang and Dennis E. Shasha. 1989. Simple Fast Algorithms for the Editing Distance Between Trees and Related Problems. *SIAM J. Comput.* 18, 6 (1989), 1245–1262. https://doi.org/10.1137/0218082

[38] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware Attention and Supervised Data Improve Slot Filling. In *Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/d17-1004

[39] Zexuan Zhong and Danqi Chen. 2021. A Frustratingly Easy Approach for Entity and Relation Extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 50–61. https://doi.org/10.18653/v1/2021.naacl-main.5

[40] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2921–2929. https://doi.org/10.1109/CVPR.2016.319

[41] Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics(volume 2: Short papers)*. 562–568. https://doi.org/10.18653/v1/p16-2034

[42] Qingfu Zhu, Wei-Nan Zhang, Ting Liu, and William Yang Wang. 2020. Counterfactual Off-Policy Training for Neural Dialogue Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 3438–3448. https://doi.org/10.18653/v1/2020.emnlp-main.276

A APPENDIX

A.1 Dataset and Platform

We use three in-domain and one out-of-domain datasets for evaluation, including **PubMed** [17][4], **TACRED** [38][5], **SemEval** [6][6], and **ACE2005** [25][7]. The data statistic and data splits are shown in Table 11.

Table 11: Data statistic and data splits. "C" and "I" denote the number of categories and instances, respectively.

Dataset	PubMed		TACRED		Semeval		ACE2005	
	C	I	C	I	C	I	C	I
Train	5	5313	42	68124	10	8000	10051	6
Validation	5	200	42	22631	-	-	2424	6
Test	5	1474	42	15509	10	2717	2050	6

In-Domain Data. Three in-domain datasets are used for two RE tasks including cross-sentence n-ary RE and sentence-level RE. For cross-sentence n-ary RE, we use the PubMed dataset which contains 4 relation labels and 1 special *none* label. For sentence-level RE task, we follow the experiment settings in Guo et al. [5] to evaluate our model on two datasets. (1) SemEval 2010 Task 8 dataset contains 9 directed relations and 1 *Other* class. (2) TACRED dataset contains 41 relation types and 1 *no_relation* class.

Out-of-Domain Data. One out-of-domain dataset is used to verify the model robustness with CAD: the ACE 2005 dataset [25] with 6 relation types and 1 special *none* class. It has 6 different domains including broadcast conversation (bc), broadcast news (bn), conversational telephone conversation (cts), newswire (nw), usenet (un), and webblogs (wl).

[4]https://github.com/Cartus/AGGCN/tree/master/PubMed
[5]https://nlp.stanford.edu/projects/tacred/
[6]Dataset link: https://github.com/Cartus/AGGCN/tree/master/semeval/dataset
[7]Dataset link: https://catalog.ldc.upenn.edu/LDC2006T06

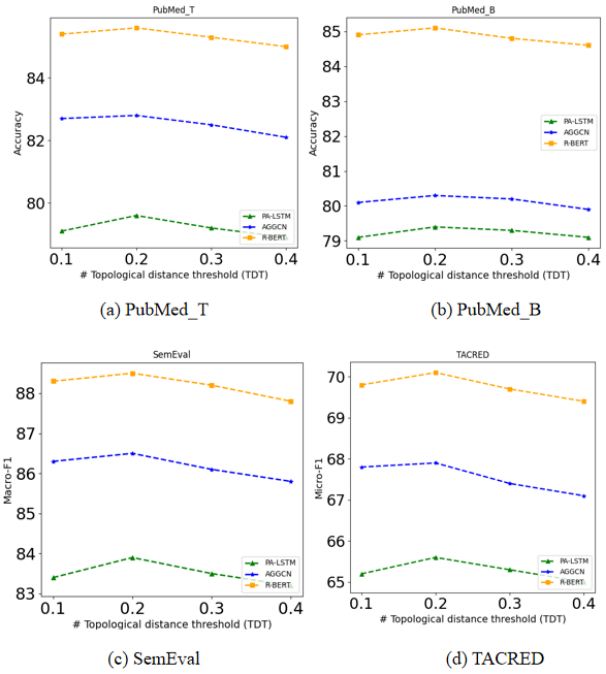


Figure 4: Impacts of TDT on PubMed, TACRED, and SemEval dataset. (_B) and (_T) denote the binary and ternary relation on PubMed.

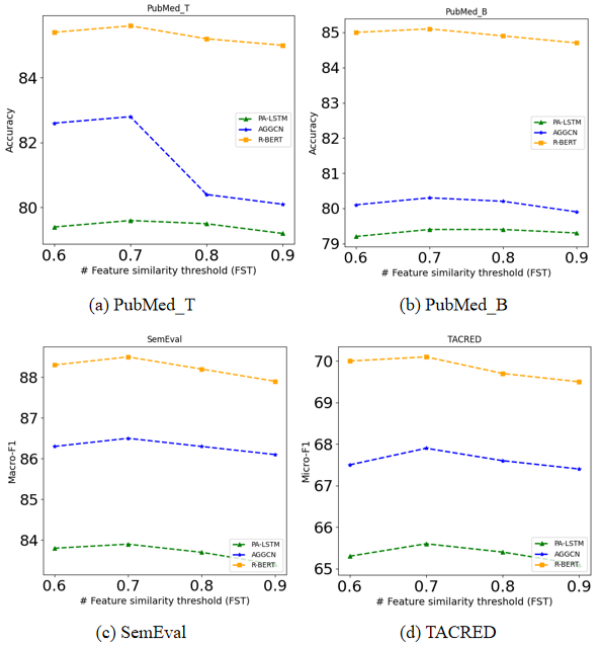
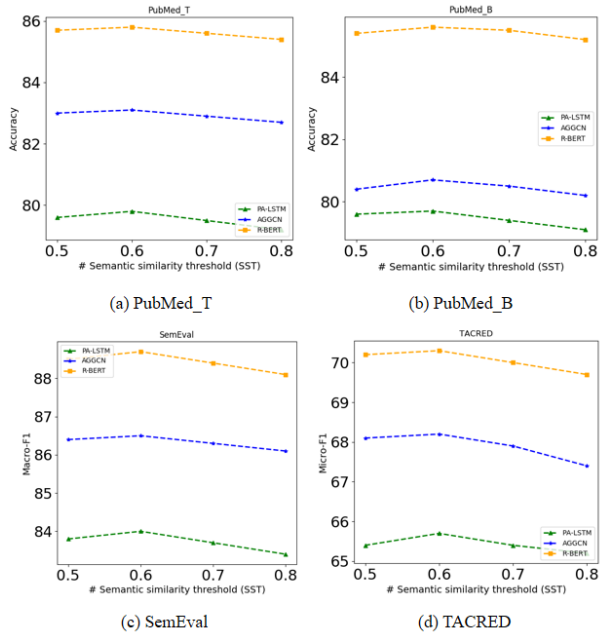


Figure 5: Impacts of FST on PubMed, TACRED, and SemEval dataset.

Platform. Our experiments are conducted on a 24GB NVIDIA 3090Ti GPU and all reported results are averaged over five runs. Our model and its variants are implemented by PyTorch.



(a) PubMed_T (b) PubMed_B

(c) SemEval (d) TACRED

Figure 6: Impacts of SST on PubMed, TACRED, and SemEval dataset.

A.2 Parameter Analysis

Our model has three key hyper-parameters including topological distance threshold (TDT), feature similarity threshold (FST), and semantics similarity threshold (SST), which determine the quality and quantity of counterfactual data generation.

We present the impacts of three key hyper-parameters on three dadasets with three backbones in Figure 4, 5, and 6, respectively.

In general, the curves for TDT, FST, and SST have similar trends in that they reach a peak first and then gradually decrease. If TDT is set too small and FST and SST are set too large, the model cannot generate enough counterfactuals to train a robust model. If TDT is set too large and FST and SST are set too small, the quality of the generated data cannot be guaranteed and the performance will also decrease.

A.3 Results for Joint Entity and Relation Extraction Tasks

We also apply our proposed method to the joint entity and relation extraction (ERE) tasks which require different backbones and different datasets. ERE includes two subtasks including named entity recognition (NER) and relation extraction (RE). Due to the space limitation, we briefly introduce the datasets and backbones and then present the results in this appendix.

Datasets. We adopt four datasets on ERE tasks, including ACE2005 [25], SciERC [12], NYT [20], and CONLL04 [22]. Their statistics are shown in Table 12.

Backbones. We adopt the latest and best ERE methods as backbones, including **PURE** [39], **PFN** [33] under joint extraction, and **UniRE** [27].

Table 12: Statistics on datasets for ERE. ϵ and R are the number of entity types and relation types. In the NYT dataset, entity type information is not annotated.

Dataset	#Sentences			ϵ	R
	Train	Dev	Test		
ACE05	10,051	2,424	2,050	7	6
SciERC	1,861	275	551	6	7
NYT	56,195	5,000	5,000	-	24
CONLL04	1,153	230	288	4	5

Table 13: Results in terms of Micro-F1 scores for ERE tasks.

Model	ACE2005	SciERC	NYT	CoNLL04
PURE$_{bb}$	64.8	36.8	91.9	72.8
+ COSY	64.8	36.9	91.9	72.9
+ RM-CT	59.3↓	33.8↓	88.1↓	69.3↓
+ GYC	62.1↓	35.2↓	90.2↓	71.1 ↓
+ CLOSS	61.3↓	34.8↓	89.1↓	69.3↓
+ CF-GAN	61.7↓	35.4↓	89.9↓	70.4↓
+ Gcc	65.2	37.2	92.1	73.1
+ **CoCo**	**65.8**	**37.7**	**92.5**	**73.3**
PFN$_{bb}$	64.6	38.4	92.4	73.6
+ COSY	64.1	38.1	92.2	73.5
+ RM-CT	60.3↓	34.3↓	88.6↓	69.6↓
+ GYC	63.3↓	36.9↓	90.5↓	72.3↓
+ CLOSS	61.5↓	34.9↓	88.9↓	67.8↓
+ CF-GAN	62.5↓	34.4↓	88.9↓	69.4↓
+ Gcc	64.8	38.7	92.6	73.8
+ **CoCo**	**65.3**	**39.2**	**92.8**	**74.0**
UniRE$_{bb}$	64.3	36.9	92.2	72.5
+ COSY	64.2↓	36.1↓	92.1↓	72.5
+ RM-CT	60.6↓	33.2↓	88.9↓	69.2↓
+ GYC	63.8↓	34.9↓	91.4↓	71.9↓
+ CLOSS	61.8↓	32.9↓	89.6↓	69.4↓
+ CF-GAN	62.2↓	33.7↓	90.4↓	70.0↓
+ Gcc	64.5	37.2	92.4	72.7
+ **CoCo**	**65.3**	**37.6**	**92.5**	**73.1**

Comparison Results for ERE Tasks. We present the results for ERE tasks. Note that we also get significant improvements on the named entity recognition (NER) task. Indeed, our CoCo can get a larger improvement on the RE task after NER. However, since this paper is focused on relation extraction (RE), we omit the comparison results for NER and only present those for RE in Table 13.

It can be seen from Table 13 that our CoCo model also significantly improves the performance of all backbones across four datasets for ERE tasks. Specifically, CoCo outperforms three backbone approaches by around 1 absolute percentage points on ACE2005 and SciERC dataset. It also improves F1 scores by 0.5-1.0 absolute percentage points on NYT and CoNLL. These results strongly demonstrate that our proposed CoCo model can consistently improve the performance of the state-of-the-art backbone methods for both the RE task or the joint ERE task.