Causal Inference-Based Debiasing Framework for Knowledge Graph Completion

Lin Ren, Yongbin Liu[✉], and Chunping Ouyang

University of South China, Hengyang, China
`yongbinliu03@gmail.com`

Abstract. The task of Knowledge Graph Completion (KGC) entails inferring missing relations and facts in a partially specified graph to discover new knowledge. However, the discrepancy in the targets between the training and inference phases might lead to in-depth bias and in-breadth bias during inference, potentially resulting in incorrect outcomes. In this work, we conduct a comprehensive analysis of these biases to determine their extent of impact. To mitigate these biases, we propose a novel debiasing framework called Causal Inference-based Debiasing Framework for KGC (CIDF) by formulating a causal graph and utilizing it for causal analysis of KGC tasks. The framework incorporates In-Depth Bias Mitigation to diminish the bias on feature representations by measuring the bias during inference, and In-Breadth Bias Mitigation to increase the distinguishability between feature representations by introducing a novel loss function. We evaluate the effectiveness of our proposed method on four benchmark datasets - WN18RR, FB15k-237, Wikidata5M-Trans, and Wikidata5M-Ind, achieving improvements of 2.5%, 0.9%, 3.2%, and 1.5% on Hit@1 respectively. Our results demonstrate that CIDF leads to significant improvements on these datasets, with more substantial gains observed in the biased settings on WN18RR achieving a 3.4% improvement in Hit@1.

Keywords: Knowledge Graph Completion · Causal Inference · Link Prediction

1 Introduction

Knowledge Graphs (KGs) are structured representations of factual knowledge, currently often represented by triples consisting of head entities, relations, and tail entities, as well as their descriptions. They are widely applied in various fields, such as question answering [9,48], dialogue systems [13,24], recommender systems [38,42], and so on. Some examples of available KGs include WordNet [23], Freebase [1], and DBpedia [18]. Building high-quality KGs often relies on human-curated structured or semi-structured data. Although many resources have been expended to refine the KGs, they remain incomplete.

Knowledge Graph Completion (KGC) is a vital task for constructing and enhancing KGs, as it involves inferring missing factual triples. Existing KGC methods generally consist of three steps: firstly, formulating a score function to measure the plausibility of triples; secondly, learning representations of entities and relations from established knowledge graphs by optimizing the scores of all factual triples; finally, using the score function to measure the plausibility of the missing triples (either a relation or entity is unknown) given the rest of the information [45], such as TransE [2], RotatE [34], HAKE [49], and ConvE [7]. In addition, due to the powerful semantic acquisition capability of Pre-training Language Models (PLMs) [8,11,31], some KGC approaches use PLMs as Knowledge Bases to leverage the semantic information from the descriptions of entities and relations such as KEPLER [39] and SimKGC [37].

However, since the training strategy for KGC tasks aims to obtain feature representations of entities and relations that fit the original knowledge graph, the process's positive and negative selection depends solely on the original knowledge graph's structure, which can often increase the correlation among factual triples' entities and relations while decreasing the correlation between those that are not within factual triples. In other words, within the knowledge graph (KG), the homophily and structural equivalence of labeled nodes and edges cause feature representations to be more sensitive to connection patterns among structurally equivalent or homophilic nodes and edges during the training process [12]. This often leads to spurious correlations in the features, creating a preference for certain connection patterns during the inference process. Additionally, the entities associated with a large number of other entities may be more likely to receive preferences in KG inference. In this study, we term the preference as **Structure Preference**, which results in the dissemination of correlation through two distinct mechanisms: in-depth diffusion and in-breadth aggregation.
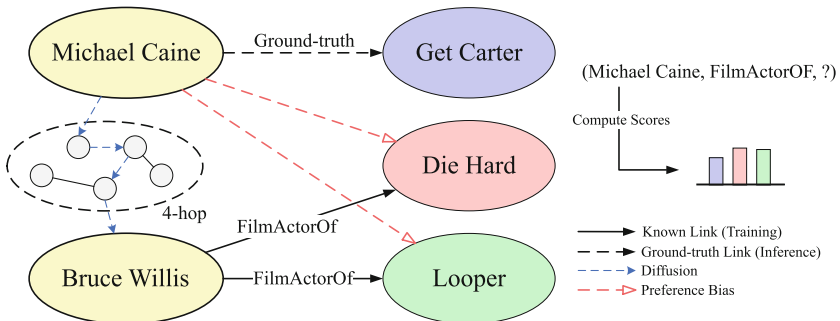

Fig. 1. A simple example from dataset FB15k-237 [35] for **in-depth bias**. *Diffusion* means the correlation diffusion along known links.

In the context of in-depth diffusion, the phenomenon may introduce a bias, which arises from the diffusion of correlation along relations between distantly

connected entities. As shown in Fig. 1, there will be a relatively high correlation between *Michael Caine* and *Bruce Willis* after training. As a result, the prediction for the triple *(Michael Caine, FilmActorOf, ?)* is inaccurately classified as *Lopper* or *Die Hard* instead of the ground-truth prediction of *Get Carter*. We refer to this phenomenon as **in-depth bias**.

In the context of in-breadth aggregation, the entities tend to aggregate information from the neighboring entities and their corresponding relations. As a consequence of the training objective, the feature representations of entities will display a higher degree of similarity if there are similar bipartite graph structures between two head entities in a certain relation and a large overlap of tail entities for this relation in the knowledge graph. This phenomenon, termed **in-breadth bias**, results in prediction biases during inference. As shown in Fig. 2, *Michael Caine* and *Joseph Gordon-Levitt* have similar topological structures, and if we want to extend the *FilmActorOf* relation for these two entities, their outcomes will exhibit a high degree of similarity, which could potentially lead to errors.
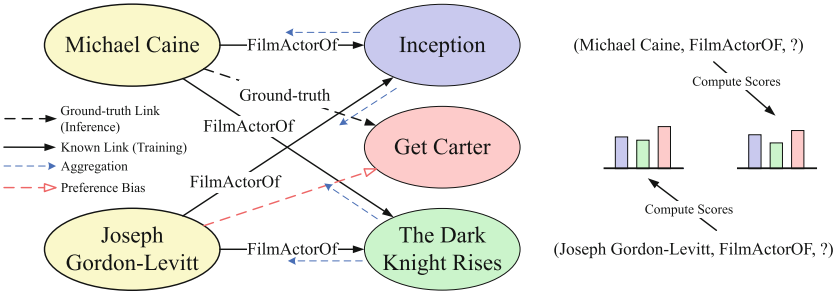

Fig. 2. A simple example from dataset FB15k-237 [35] for **in-breadth bias**. *Aggregation* means the correlation aggregation through known links.

In this paper, inspired by causal inference [28], we formulate a causal graph to model the KGC task and introduce corresponding causal methods to mitigate in-depth bias and in-breadth bias, called **Causal Inference-based Debiasing Framework for KGC (CIDF)**. For in-depth bias, inspired by debiasing methods of counterfactual analysis [26,30,40], we leverage the trained representations suffering from the bias to compute individual in-depth bias for each representation and global in-depth bias for all representations, using them to obtain the debiased representations. We call this method In-Depth Bias Mitigation (DBM). Additionally, for addressing in-breadth bias, we introduce a novel loss function in the training phase to reduce the similarity among feature representations, enhancing the distinguishability of different entities and relations. We call this method In-Breadth Bias Mitigation (BBM).

We evaluate the effectiveness of our proposed CIDF on four commonly used datasets, namely WN18RR, FB15k-237, Wikidata5M-Trans, and Wikidata5M-Ind, through a series of experiments. By applying CIDF to SimKGC [37], we

observe improved performance across all evaluation metrics, including MRR, Hit@1, Hit@3, and Hit@10, with improvements of 2.5%, 0.9%, 3.2%, and 1.5%, respectively, over the Hit@1 baseline on the aforementioned datasets. Notably, our framework shows greater improvement in a biased setting.

2   Related Work

2.1   Knowledge Graph Completion

Knowledge graph completion (KGC) involves automatically inferring missing or uncertain facts in a given knowledge graph. Different models have been proposed to tackle this problem, such as translational models like TransE [2], TransH [41], and RotatE [34], which interpret relations as translation or rotation operations, and tensor decompositional models like RESCAL [27], DistMult [43], and ComplEx [14,17,36], which treat KGC as a tensor factorization problem. More recent approaches, such as LP-BERT [19], KEPLER [39], and SimKGC [37], attempt to exploit textual information. However, the majority of the above models focus on acquiring feature representations to fit original KGs and fail to consider in-depth bias and in-breadth bias.

2.2   Bias in Knowledge Graph

Entity and relation representations in a knowledge graph are obtained by statistically fitting existing facts and summarizing their distributional characteristics [16]. Consequently, the representation may be influenced by biases introduced through the use of statistical methods. Several works have emerged with a focus on mitigating biases in knowledge graphs. Some of these works include [20,33,50], which concentrate on mitigating the degree bias; [3,10], which address gender bias reduction; [21,32], which consider sensitive information in KGs; and [15], which aims to detect biases automatically. In this work, we concentrate on in-depth bias and in-breadth bias, and propose a method to mitigate them.

2.3   Causal Inference

Causal inference is a statistical technique that enables the identification of causal relationships between variables in complex systems. This method has found extensive applications in various fields, including semantic segmentation [46], video grounding [25], stable learning [47], text classification [30], medical Q&A [44], and information extraction [26,40]. Recently, causal inference has also been employed in KGC, such as KGCF [4] which utilizes causal inference for data augmentation on NBFNet [52] to achieve improvement. Additionally, GEAR [21] and CFLP [51] have introduced counterfactual inference in KGC, demonstrating substantial enhancements in performance. In this work, we employ causal inference to mitigate in-depth and in-breadth biases present in KGC tasks.

3 Preliminaries

This section intends to offer a comprehensive overview of the notation employed, the inference process of knowledge graph completion methods, and the specific setting selected for this study.

3.1 Notation

We denote a Knowledge Graph as $G = (E, R, T)$, where E represents the set of entities, R refers to the set of relations, and T is the set of triples that constitute the knowledge graph. A triple consists of a head entity, a relation, and a tail entity, $(head\ entity, relation, tail\ entity)$, where the head entity is the initiator of the relation and the tail entity is the recipient of the relation, in which the order of the two entities cannot be reversed. To simplify the description, h, r, and t are used to represent *head entity*, *relation*, and *tail entity*, respectively, with (h, r, t) denoting a triple.

The evaluation protocol for entity ranking has gained widespread adoption in the field of KGC, which aims to rank all entities based on their relevance to a given entity and relation pair. Specifically, KGC tasks include tail entity prediction, represented as $(h, r, ?)$, and head entity prediction, represented as $(?, r, t)$. To reformulate $(?, r, t)$ as $(t, r^{-1}, ?)$, where r^{-1} is the inverse relation of r, we can focus solely on tail entity prediction task [22].

3.2 Inference Process of KGC Methods

KGC methods propose various scoring functions to optimize entity and relation representations. These methods strive to learn representations that capture the underlying semantic information in the knowledge graph, facilitating more accurate predictions of missing factual triples. To offer a thorough understanding of these methods, we present an overview of several KGC approaches below.

TransE [2] is a traditional translational method for representation learning in knowledge representation, which considers relation as translational operations. To achieve this, the optimization target in the training phase is the minimization of the distance between the expected tail entity and the translated head entity after applying the relation representation, expecting $h + r = t$ established finally if (h, r, t) is a missing factual triple. In RotatE [34], relations between entities are modeled as rotation operations within complex space, with the primary goal of satisfying the equation $h \odot r = t$ where $\odot$ denotes the Hadamard product. SimKGC [37] employs the BERT [8] model to extract the fusion feature of the head entity and relation, denoted as f_{hr}, and the representation of the tail entity, denoted as f_t. The objective of SimKGC is to maximize the similarity between f_{hr} and f_t. Table. 1 presents a comprehensive overview of the score functions utilized by the KGC methods outlined above.

To guarantee a clear causal analysis and modeling process, we opt for the SimKGC framework as the fundamental setting due to its simplicity and effectiveness. Employing this framework as a foundation allows us to build upon it for comprehensive analysis and modeling.

Table 1. Score functions for KGC methods. Where $cossim(\cdot)$ means cosine similarity function, and $\uparrow$ indicates that higher values are better.

Methods	Score Function $\uparrow$
TransE [2]	$-||h + r - t||_{l_1/l_2}$
RotatE [34]	$-||h \odot r - t||$
SimKGC [37]	$cossim(f_{hr}, f_t)$

4   Methodology

4.1   Bias Analysis

The majority of KGC techniques rely on crafted scoring functions to act as optimization targets for learning entity and relation representations. Typically, these methods partition the data into two distinct groups: positive samples, which correspond to triples that present in the knowledge graph, and negative samples, which correspond to non-existent triples, and then they seek to maximize the aggregate score of all positive samples while minimizing the aggregate score of all negative samples, as shown in Eq. 1. This training process yields a learned representation of entities and relations that can then be used to determine the missing factual triples in the KG. By optimizing the scoring functions and learning robust entity and relation representations through machine learning, KGC methods have achieved significant success in accurately and efficiently inferring missing information to complement KGs.

$$\underset{Rep_E, Rep_R}{argmax} \sum_{(h,r,t) \in T} f(e_h, e_r, e_t) \; - \sum_{(h,r,t) \notin T} f(e_h, e_r, e_t) \tag{1}$$

Where $f(\cdot)$ denotes the score function, Rep_E and Rep_R represent representation sets of E and R respectively; T denotes factual triples of the training set; $e_h, e_t \in Rep_E$ are the representations of entities h and t; $e_r \in Rep_R$ is the representation of relation r.

However, optimizing the aforementioned objective may cause issues in KGC tasks. Specifically, the absent triples in the original knowledge graph are treated as negative samples, leading to their scores being diminished in the training process. This not only affects non-factual triples but also impacts missing factual triples. Additionally, the entities will be correlated during training if there are paths connecting them in the KG. Consequently, during the inference phase, given a task $(h, r, ?)$, the entities connected with h in the original KG will more likely receive high scores, while ineligible entities will receive relatively lower scores, resulting in an unfair issue as shown in Fig. 1. Although addressing this bias may be hard for the KGC methods which solely rely on the structure of KGs, such as TransE [2], it can be measured and mitigated in methods that leverage descriptions of entities and relations as additional information.

Furthermore, it's noteworthy that the aforementioned phenomenon may also present an additional issue when dealing with entities that exhibit similar topo-

logical structures within a KG. Consequently, given two tasks with the form $(h_1, r, ?)$ and $(h_2, r, ?)$, where h_1 is topologically similar to h_2, it is reasonable to expect that the outcomes of these tasks will exhibit significant similarity. Figure 2 shows a simple example of this issue, we call it **in-breadth bias** in this paper.
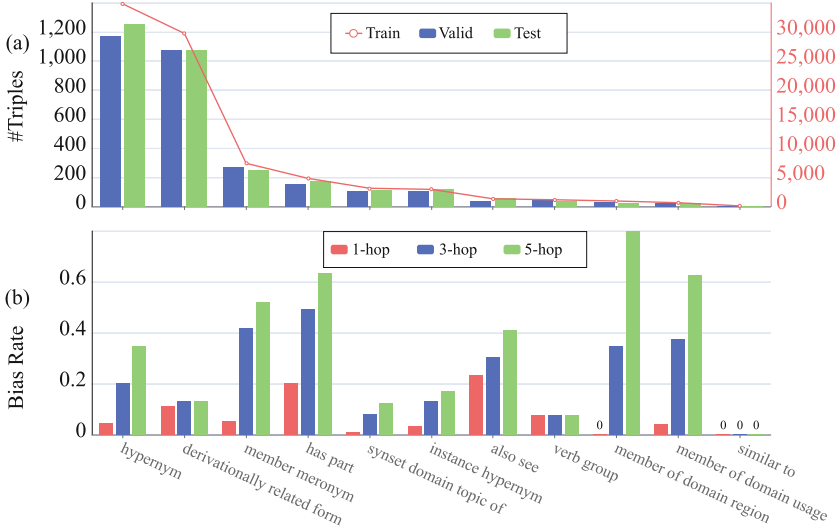

Fig. 3. The statistic of WN18RR where the x-axis represents relation names, noting that the x-axis for (a) and (b) is the same. (a) Quantity statistics for each relation in the dataset where the y-axis means the number of triples in train, validation, and test sets respectively. The red y-axis corresponds to the value of *Train*. (b) Bias statistic for each relation in the dataset where the y-axis means *bias rate* in different settings. (Color figure online)

In order to assess the degree of bias presented in KGs, we conduct a statistical analysis on datasets commonly used in KGC tasks. Specifically, we focus on the WN18RR dataset, which has been widely employed in KGC research with relatively few relations, as shown in Fig. 3. Firstly, we conduct an analysis of the quantity statistics and discovered a significant long-tail distribution among the relations in the dataset. In particular, as shown in Fig. 3(a), we observe that while the *hypernym* relation appears in tens of thousands of triples in the training set, the *similar to* relation appears only in tens of triples, and the proportions of relation numbers are similar in the validation and test sets. Moreover, to measure the degree of in-depth bias in KG datasets, we employ a function to quantify the *bias rate* of each relation as follows:

$$BiasCount_{r_i}^{hop} = \sum_{(h,r,t) \in Res_t} \mathbb{I}((h,r,t) \notin FT \wedge C(h,t,hop) \wedge r = r_i) \quad (2)$$

$$SampleCount_{r_i} = \sum_{(h,r,t) \in Res_t} \mathbb{I}((h,r,t) \notin FT \wedge r = r_i) \tag{3}$$

$$BiasRate_{r_i}^{hop} = BiasCount_{r_i}^{hop}/SampleCount_{r_i} \tag{4}$$

where $hop \in \{1,3,5\}$ denotes whether two entities are $\{1,3,5\}$-hop neighbors; $C(h,r,k)$ denotes an indicator function that the value is 1 if h and t are k-hop neighbors, 0 if not; $\mathbb{I}(\cdot)$ represents indicator function that the value is 1 if the condition holds, 0 if not; $r_i \in R$ denotes a specific relation; FT denotes the factual triples of the testing set; Res_t means the prediction results of the testing set from SimKGC framework; r^{-1} is viewed as the same relation type as r. In a nutshell, *bias rate* refers to the ratio of prediction tasks that may be impacted by in-depth bias during inference.

As depicted in Fig. 3, a significant proportion of relations in the WN18RR datasets may be affected by the issue of in-depth bias. This issue is especially pronounced for specific relations, such as *member meronym* and *has part* in the 3-hop setting. In the 5-hop context, these bias rates exhibit a greater increase.

4.2   Causal Analysis

As a foundation of causal analysis, we select SimKGC's framework [37], which has a relatively simple structure in inference, referring to $\underline{hr}$ as a fuse representation for h and r which is utilized to measure the rank of all entities. As shown in Fig. 4, we formulate a **causal graph** [28,29] for KGC tasks. The causal graph is denoted as a directed acyclic graph which consists of a variable set and a directed edge set, which can provide a comprehensive visualization of causal relationships.

In the causal graph, we consider the dependencies of feature representations of hr and t on both the training set (KG), which comprises the factual triples (topological structures) of the original KG, and the textual descriptions (S_{hr} and S_t), which define as the semantic information of hr and t. Furthermore, the score calculation (SC) of a given pair of hr and t is accomplished by utilizing their feature representations (F_{hr} and F_t).

The causal graph demonstrates that feature representation learning is affected not only by the structure of the knowledge but also by the semantic information of entities and relations. The biases that arise from in-depth and in-breadth both relate to the path $F_{hr} \leftarrow KG \rightarrow F_t$. Consequently, the mitigation of the above biases can be achieved by exerting influence over the aforementioned path.

4.3   In-Depth Bias Mitigation

The cause of in-depth bias in KGC tasks lies in the structural equivalence preference during training, where the correlation between entities and relations will gradually diffuse along links of KGs. Consequently, entities that are distant from each other may result in incorrect outcomes due to high correlations. This bias can be reduced by modifying the feature representations of hr according to the structure preference during inference.
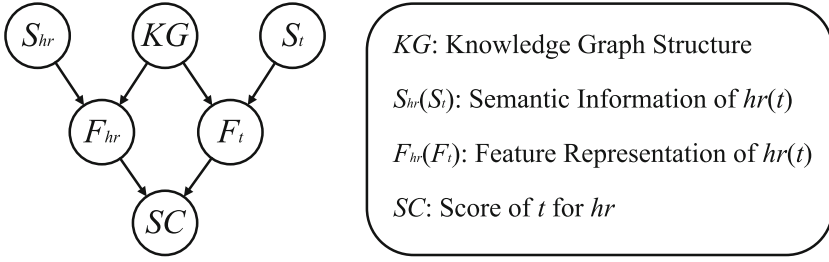
Fig. 4. The causal graph of knowledge graph completion. KG can be seen as the training set; S_{hr} and S_t are the text description of corresponding entities and relations. It should be noted that unobserved biases which can not be observed and interfered with are ignored in this figure.

Inspired by **Counterfactual Analysis** [26,28,30,40], it incorporates a counterfactual scenario into the debiasing process for a given task. The construction of the scenario serves as a guiding factor in mitigating bias in the inference. To this end, we construct a counterfactual inference scenario for the KGC task and present a factual inference scenario, the original inference scenario of KGC tasks, for comparison:

- **Factual Inference:** *Given $(h, r, ?)$, what will the missing entity be if seeing the original KG?*
- **Counterfactual Inference:** *Given $(h, r, ?)$, what will the missing entity be if seeing the original KG and knowing that in-depth bias may cause incorrect outcomes?*

In order to enable the hr to effectively "recognize" the correlated tail entities (t) caused by the structure preference, we leverage the phenomenon that the triples consisting of correlated entities and relations tend to exhibit relatively high scores while other triples demonstrate relatively low scores. Through leveraging the aforementioned phenomenon, it becomes possible to measure and quantify the extent of in-depth bias that exists for hr and t during the inference process. Following this way, we introduce two measurable in-depth biases, namely, **individual in-depth bias** and **global in-depth bias**, and their description are as follows:

- **Individual In-Depth Bias (IDB):** The bias caused by the path $KG \rightarrow F_{hr}$ indicates the overall feature of the set of t that can be correlated with hr during training. It is different for each hr.
- **Global In-Depth Bias (GDB):** The bias caused by path $KG \rightarrow F_t$ indicates the overall bias of the KG that the entities correlated with more other entities are predisposed to attain higher scores. This bias has uniqueness for each dataset.

To assess the **individual in-depth bias**, we employ hr to assign weights to all entities, and then calculate the weighted average of these weights to serve as

the bias feature:

$$F_{IDB}^{hr} = norm(\sigma(\sum_E^t w_t^{hr} * F_t))$$

$$w_t^{hr} = \exp(\hat{w}_t^{hr})/\sum_E^i \exp(\hat{w}_i^{hr})$$

$$\hat{w}_t^{hr} = cossim(F_{hr}, F_t)$$

(5)

where F_{IDB}^{hr} means the individual in-depth bias of hr; $cossim(\cdot)$ means cosine similarity function; i and t means a entity in E; $\sigma(\cdot)$ means nonlinear activation function $tanh$; $norm(\cdot)$ denotes the L_2 normalization function.

In order to access the **global in-depth bias**, the bias feature is represented by the average feature of all entities which represents the overall preference of the dataset:

$$F_{GDB} = \frac{1}{N} \sum_E^t F_t$$

(6)

where F_{GDB} means the global in-depth bias of the KG and N means the number of entities.

Finally, F_{IDB}^{hr} and F_{GDB} are used to compute the debiasing result of F_{hr}:

$$F_{hr}^{db} = F_{hr} - \lambda_{IDB} F_{IDB}^{hr} - \lambda_{GDB} F_{GDB}$$

(7)

where F_{hr}^{db} is the debiasing feature of hr after in-depth bias mitigation; λ_{IDB} and λ_{GDB} are hyperparameters. F_{hr}^{db}, replacing F_{hr}, is used to predict the missing entity.

4.4 In-Breadth Bias Mitigation

The in-breadth bias obtained during inference can be attributed to the training strategy adopted. In a certain relation, there are similar bipartite graph structures and a large overlap of tail entities. Specifically, those hr or t with similar topological structures are likely to acquire similar feature representations.

The bias results in the path $F_{hr} \leftarrow KG \rightarrow F_t$, as illustrated by the causal graph Fig. 4. The main cause of this bias lies in the fact that the training strategy employed does not sufficiently account for the potential overlap in features between hr or t that are structurally similar.

To reduce this bias, **causal intervention** [28] is used to block paths $KG \rightarrow F_{hr}$ and $KG \rightarrow F_t$ by keeping F_{hr} and F_t as constants during training. But it is impossible in KGC tasks due to the necessity of learning feature representations through the structure of KGs. Therefore, we reduce the restriction of the aforementioned method and mitigate the impact from $F_{hr} \leftarrow KG \rightarrow F_t$ by respectively lowering the similarities between representations of hr and between those of t during training. In practice, we introduce a novel loss function that quantifies the level of similarity among feature representations within a batch, and minimize it to reduce **in-breadth bias (BB)** during the training phase as follows:

$$L_{BB} = \frac{1}{2N} \sum_N^i \sum_N^j (cossim(F_{hr}^i, F_{hr}^j) + cossim(F_t^i, F_t^j))$$

(8)

where N means the sample number of one batch during training; F_{hr}^i and F_t^i mean the $i-th$ F_{hr} and F_t respectively in the batch; $cossim(\cdot)$ means cosine similarity function. L_{BB} is used as an extra loss in the original training phase.

5   Experiment

5.1   Datasets

We conduct experiments to evaluate the performance of our method **CIDF** including **In-Depth Bias Mitigation (DBM)** and **In-Breadth Bias Mitigation (BBM)** on three widely used KGC datasets: WN18RR [7], FB15k-237 [35], and Wikidata5M[1] [39]. The data statistics of each dataset are shown in Table. 2. WN18RR and FB15k-237 are revised by WN18 and FB15k datasets [2], respectively, which suffer from test set leakage. For the Wikidata5M dataset, there exist two distinct settings available, namely transductive and inductive. In the transductive setting, the set of entities is identical between the training and inference phases, whereas, in the inductive setting, entities encountered during training will not appear during inference.

Table 2. Statistics of datasets we use in experiments. Wikidata5M-Trans and Wikidata5M-Ind denote the transductive and inductive settings of Wikidata dataset respectively.

Dataset	#Entity	#Relation	#Taining	#Valid	#Test
WN18RR	40,943	11	86,835	3,304	3,314
FB15k-237	14,541	237	272,115	17,535	20,466
Wikidata5M-Trans	4,594,485	822	20,614,279	5,163	5,163
Wikidata5M-Ind	4,579,609	822	20,496,514	6,699	6,894

5.2   Evaluation Metrics

Following previous KGC works, we evaluate the performance of our proposed method with the entity ranking task. In practice, for each task $(h, r, ?)$ during test phase, it is required to predict the ranks of all entities as the result of t given h and r, which is similar for task $(t, r^{-1}, ?)$. For evaluating the overall performance of datasets, we use four evaluation metrics as follows: Mean Reciprocal Rank (MMR), and Hit@k ($k \in \{1, 3, 10\}$), where MRR is determined by the average reciprocal rank of all test triples and Hit@k is calculated from the proportion of correct entities ranked among the top-k. Additionally, in main comparative experiments, MRR and Hit@k are computed following the *filtered setting* [2] which excludes the factual triples when ranking.

---

[1] https://deepgraphlearning.github.io/project/wikidata5m.

5.3 Baselines

For comparison, we select a series of strong baselines as follows:

- KGC methods that integrate head entity and relation as a whole in score computing, including SimKGC [37], Hitter [5], and LP-BERT [19].
- KGC methods that treat relations as operations in score computing, including TransE [2], RotatE [34], DistMult [43], and KEPLER [39].

5.4 Hyperparameters

Given that our proposed method, CIDF, is primarily applied within the SimKGC framework, we maintain the existing hyperparameters setting within SimKGC, which aims to ensure that the comparison is consistent and fair. The two encoders for hr and t are both the pre-training language model called *bert-base-uncased* [8]. The descriptions in both hr and t are restricted to a maximum token number of 50. Any portion of the descriptions that exceeds this limit will be truncated. Within the hyperparameters of CIDF, we conduct grid searches on λ_{IDB} and λ_{GDB} with ranges $[0.1, 1]$ at 0.1 intervals. The training epochs for datasets WN18RR, FB15k-237, and Wikidata5M are 50, 10, and 1 respectively. In the training phase, the models are run on 4 V100 GPU (32G) with batch size 1024.

Table 3. Main results for datasets WN18RR and FB15k-237. "CIDF" refers to our proposed method Causal Inference based KGC Debiasing Framework;"-BB", "-GDB" and "-IDB" refer to the result after mitigating In-Breadth Bias, Global In-Depth Bias, and Individual In-Depth Bias respectively. SimKGC is the setting SimKGC$_{IB+PB+SN}$ [37].

Method	WN18RR				FB15k-237			
	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
TransE(2013) [2]	22.2	1.2	40.1	52.9	32.9	23.0	36.8	52.7
DisMult(2015) [43]	44.0	39.6	45.2	53.6	30.1	22.1	33.8	48.3
RotatE(2019) [34]	47.7	42.9	49.4	57.2	33.5	23.7	37.3	53.1
LP-BERT(2022) [19]	48.2	34.3	**56.3**	**75.2**	31.0	22.3	33.6	49.0
Hitter(2021) [5]	**49.9**	**45.7**	51.3	58.7	**37.2**	**27.8**	**40.8**	**55.8**
CIDF								
SimKGC(2022) [37]	66.5	58.6	71.5	80.2	33.6	24.9	36.2	51.1
-BB	67.8	60.6	72.0	80.3	33.8	25.3	36.6	51.4
-BB-GDB	67.9	60.9	72.1	80.4	34.4	25.6	37.1	52.3
-BB-GDB-IDB	**68.1**	**61.2**	**72.2**	**80.4**	**34.6**	**25.8**	**37.4**	**52.7**

5.5 Main Results

The results of the KGC tasks conducted on WN18RR and FB15k-237 datasets are presented in Table. 3. It indicates that the CIDF applied to the SimKGC

framework enhances all evaluation metrics across both two datasets. In comparison to SimKGC$_{\text{IB+PB+SN}}$, applying CIDF for WN18RR shows improvements of 1.6%, 2.6%, 0.7%, and 0.2% in MRR, Hit@1, Hit@3, and Hit@10, correspondingly. Similarly, for FB15k-237, applying CIDF offers enhancements of 1.0%, 0.9%, 1.2%, and 1.6% in MRR, Hit@1, Hit@3, and Hit@10, respectively.

Table 4. Main results for dataset Wikidata5M.

Method	Wikidata5M-Trans				Wikidata5M-Ind			
	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
TransE(2013) [2]	25.3	17.0	31.1	39.2	-	-	-	-
RotatE(2019) [34]	**29.0**	**23.4**	**32.2**	**39.0**	-	-	-	-
KEPLER(2021) [39]	21.0	17.3	22.4	27.7	40.2	22.2	51.4	73.0
BLP-ComplEx(2021) [6]	-	-	-	-	48.9	26.2	**66.4**	**87.7**
BLP-SimplE(2021) [6]	-	-	-	-	**49.3**	**28.9**	63.9	86.6
CIDF								
SimKGC(2022) [37]	35.8	31.3	37.6	44.1	71.4	60.9	78.5	91.7
-BB	36.8	32.2	38.2	44.7	**72.6**	**62.4**	**79.8**	**91.7**
-BB-GDB	37.4	32.9	39.0	45.5	-	-	-	-
-BB-GDB-IDB	**38.7**	**34.5**	**40.2**	**46.7**	-	-	-	-

Furthermore, we present the performance results of our proposed method on the Wikidata5M dataset, presented in both the transductive and inductive settings, as depicted in Table. 4. In the Wikidata5M-Trans setting, our method displays a notable increase of 2.9%, 3.2%, 2.6%, and 2.6% improvements in MRR, Hit@1, Hit@3, and Hit@10, respectively. On the other hand, in the Wikidata5M-Ind setting, since the entities utilized during the training and inference phases are mutually exclusive, we only conduct in-breadth bias mitigation (-BB) in this setting and achieve enhancements in MRR, Hit@1, and Hit@3 by 1.2%, 1.5%, and 1.3%, respectively.

In summary, the experimental results presented above indicate that our proposed method, Causal Inference based KGC Debiasing Framework (CIDF), effectively enhances the performance of KGC tasks.

6   Analysis

In this section, we aim to further illuminate the effectiveness of our proposed method by carrying out a series of analytical experiments from various perspectives.

6.1   Fine-Grained Analysis

We conduct a series of fine-grained analytical experiments on WN18RR, a knowledge graph with a relatively smaller number of relations with well-defined label descriptions. We calculate the performance metrics for each relation specifically, using the mean reciprocal rank (MRR), as our primary evaluation metric.

Table 5. The fine-grained results of WN18RR basing SimKGC framework. **"Bias Rate"** is computed by Eq. 4; **"Prop."** denotes the proportion for each relation in the training set; **"overall"** denotes measuring the results for all samples while **"k-hop"** ($k \in \{3,5\}$) denotes measuring the results only for the samples with k-hop in-depth bias; **"w/o."** and **"w."** denote the results without and with CIDF respectively; **"avg. Hit@1"** denotes the average Hit@1 for all relations; ↑ (↓) means the result is better when the value is higher (lower). Relation abbreviation: "derivational .. form" = "derivationally related form", "synset .. of" = "synset domain topic of", "member .. region" = "member of domain region", and "member .. usage" = "member of domain usage". ♢♡♣♠ are the markers for facilitating locating the specific part of this table in the paper.

| Relation | Prop.(%) | MRR ↑ | | | | Bias Rate ↓ | | | |
| | | overall♢ | | biased♡ | | 3-hop♣ | | 5-hop♠ | |
		w/o.	w	w/o.	w	w/o.	w	w/o.	w
hypernym	39.9	48.9	**50.0**	37.2	**38.1**	20.2	**18.8**	34.8	**32.6**
derivationally .. form	34.3	90.3	**93.2**	89.4	**92.3**	13.2	**10.8**	13.6	**11.3**
member meronym	8.1	58.9	**62.8**	52.0	**56.2**	41.9	**38.7**	52.2	**48.6**
has part	5.5	45.9	**49.0**	41.0	**48.2**	49.4	**47.1**	63.4	**61.6**
synset .. of	3.9	61.7	**63.2**	36.1	**40.6**	7.9	**6.1**	12.3	**9.6**
instance hypernym	3.6	66.8	**71.4**	55.9	**63.2**	13.1	**9.0**	17.2	**12.3**
also see	1.8	**67.7**	64.5	**69.9**	65.9	30.4	**26.8**	41.1	**37.5**
verb group	1.2	92.8	**96.5**	84.2	**88.0**	7.7	**5.1**	7.7	**5.1**
member .. region	0.8	46.6	**48.7**	17.6	**22.4**	34.6	**30.8**	80.8	**76.9**
member .. usage	0.8	53.2	**60.7**	32.4	**35.6**	37.5	**33.3**	62.5	62.5
similar to	0.1	100.0	100.0	100.0	100.0	0.0	0.0	0.0	0.0
avg	-	66.6	**69.1**	56.0	**59.1**	23.3	**20.6**	35.0	**32.6**
avg. Hit@1	-	60.0	**63.4**	47.7	**52.1**	-	-	-	-

Conventional Results. In the filtered setting, our analysis, as presented in Table. 5 ♢, demonstrates that after applying our proposed method CIDF, nearly all of the relations exhibit performance improvements. More specifically, on average, we observe an enhancement of 2.5% and 3.4% for MRR and Hit@1 respectively, which underscores the effectiveness of CIDF. Furthermore, when considering the symmetric relation *also see*, whereby $(A, also\ see, B)$ and $(B, also\ see, A)$ are both factual triples when one of them holds, there is a notable decrease of 3.1% in MRR that may be attributed to the effects of DBM. To further investigate this performance drop, we conduct a supplementary experiment consisting of solely using BBM of CIDF without DBM. As a result, we observe a significant increase in the MRR of *alse see* relation, which improves from 64.5% to 67.8% (+3.2%).

Biased Results. In order to assess the efficacy of our proposed method in tacking in-depth bias, we conduct performance measurements exclusively on the samples which are potentially susceptible to this bias. Table. 5 ♡ shows the results in the biased setting on the performance of relations. In the biased setting, we only evaluate the *hr* that exists in both the training and inference phases. By

applying CIDF, there are improvements of 3.2% and 4.4% on average MRR and Hit@1 respectively, which means CIDF is more effective in the biased setting.

Bias Rate Results. To assess the efficacy of CIDF in bias mitigation, bias rates for individual relations in 3-hop and 5-hop settings are measured. The results, presented in Table. 5 ♣ and ♠, demonstrate a reduction in bias rates for most relations by applying CIDF. Specifically, the average bias rates decrease by 2.7% and 2.4% in the respective hop settings.
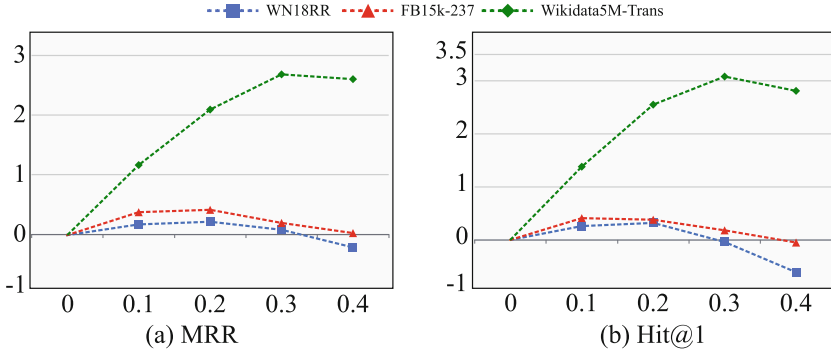

Fig. 5. Detail results of MRR and Hit@1. The x-axis means the value of λ_{IDB}; the y-axis represents the difference in performance compared with $\lambda_{IDB} = 0$.

6.2   Effect of In-Depth Bias Mitigation

For the better adaptation of DBM to different datasets, we use two hyperparameters to control the degree of bias mitigation (Eq. 5). Specifically, we conduct beam searches for λ_{IDB} and λ_{GDB} to gain the optimal setting for different datasets as depicted in Eq. 5. To effectively determine the optimal settings for various datasets, we employ beam searches for λ_{IDB} and λ_{GDB} with ranges $[0.1, 1]$ at 0.1 intervals. As shown in Fig. 5, the optimal λ_{IDB} is 0.1 for WN18RR and FB15k-237 datasets and is 0.3 for Wikidata5M-Trans. For λ_{GDB}, the optimal values of all datasets used in this paper are 1.0.

7   Conclusion and Future Work

In this paper, we present a bias analysis of Knowledge Graph Completion tasks and identify two biases, in-depth and in-breadth, during the training phase that may lead to erroneous outcomes during inference. To mitigate these biases, we conduct a causal analysis and formulate a causal graph for KGC tasks. Building on this, we propose a novel debiasing framework, the Causal Inference-biased

KGC Debiasing Framework, which incorporates In-Depth Bias Mitigation and In-Breadth Bias Mitigation. Applying CIDF results in significant improvements on three benchmark datasets, namely WN18RR, FB15k-237, and Wikidata5M, particularly in the biased setting. In the future, we intend to conduct a more comprehensive analysis of KGC task biases and develop a general causal debiasing framework applicable to various KGC methods directly.

Supplemental Material Statement: Source code, datasets and results are available at https://github.com/HomuraT/CIDF.

References

1. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1247–1250. SIGMOD 2008, Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1376616.1376746

2. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 26. Curran Associates, Inc. (2013). https://proceedings.neurips.cc/paper_files/paper/2013/file/1cecc7a77928ca 8133fa24680a88d2f9-Paper.pdf

3. Bourli, S., Pitoura, E.: Bias in knowledge graph embeddings. In: 2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), pp. 6–10 (2020). https://doi.org/10.1109/ASONAM49781.2020. 9381459

4. Chang, H., Cai, J., Li, J.: Knowledge graph completion with counterfactual augmentation. In: Proceedings of the ACM Web Conference 2023, pp. 2611–2620. WWW 2023, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3543507.3583401

5. Chen, S., Liu, X., Gao, J., Jiao, J., Zhang, R., Ji, Y.: HittER: hierarchical transformers for knowledge graph embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 10395–10407. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, November 2021. https://doi.org/10.18653/v1/2021.emnlp-main.812

6. Daza, D., Cochez, M., Groth, P.: Inductive entity representations from text via link prediction. In: Proceedings of the Web Conference 2021, pp. 798–808. WWW 2021, Association for Computing Machinery, New York, NY, USA (2021). https:// doi.org/10.1145/3442381.3450141

7. Dettmers, T., Minervini, P., Stenetorp, P., Riedel, S.: Convolutional 2d knowledge graph embeddings. In: McIlraith, S.A., Weinberger, K.Q. (eds.) Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-2018), the 30th innovative Applications of Artificial Intelligence (IAAI-2018), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-2018), New Orleans, Louisiana, USA, February 2–7, 2018, pp. 1811–1818. AAAI Press (2018). https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17366

8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, June 2019. https://doi.org/10.18653/v1/N19-1423

9. Ding, M., Zhou, C., Chen, Q., Yang, H., Tang, J.: Cognitive graph for multi-hop reading comprehension at scale. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2694–2703. Association for Computational Linguistics, Florence, Italy, July 2019. https://doi.org/10.18653/v1/P19-1259

10. Du, Y., Zheng, Q., Wu, Y., Lan, M., Yang, Y., Ma, M.: Understanding gender bias in knowledge base embeddings. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1381–1395. Association for Computational Linguistics, Dublin, Ireland, May 2022. https://doi.org/10.18653/v1/2022.acl-long.98

11. Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., Tang, J.: GLM: general language model pretraining with autoregressive blank infilling. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 320–335. Association for Computational Linguistics, Dublin, Ireland, May 2022. https://doi.org/10.18653/v1/2022.acl-long.26

12. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864 (2016)

13. Guo, D., Tang, D., Duan, N., Zhou, M., Yin, J.: Dialog-to-action: conversational question answering over a large-scale knowledge base. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31. Curran Associates, Inc. (2018). https://proceedings.neurips.cc/paper_files/paper/2018/file/d63fbf8c3173730f82b150c5ef38b8ff-Paper.pdf

14. Jia, T., Yang, Y., Lu, X., Zhu, Q., Yang, K., Zhou, X.: Link prediction based on tensor decomposition for the knowledge graph of COVID-19 antiviral drug. Data Intell. **4**(1), 134–148 (2022). https://doi.org/10.1162/dint_a_00117

15. Keidar, D., Zhong, M., Zhang, C., Shrestha, Y.R., Paudel, B.: Towards automatic bias detection in knowledge graphs. In: Findings of the Association for Computational Linguistics: EMNLP 2021. pp. 3804–3811. Association for Computational Linguistics, Punta Cana, Dominican Republic, November 2021. https://doi.org/10.18653/v1/2021.findings-emnlp.321

16. Kraft, A., Usbeck, R.: The lifecycle of "facts": a survey of social bias in knowledge graphs. In: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 639–652. Association for Computational Linguistics, Online only, November 2022. https://aclanthology.org/2022.aacl-main.49

17. Lacroix, T., Usunier, N., Obozinski, G.: Canonical tensor decomposition for knowledge base completion. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018. Proceedings of Machine Learning Research, vol. 80, pp. 2869–2878. PMLR (2018). http://proceedings.mlr.press/v80/lacroix18a.html

18. Lehmann, J., et al.: DBPedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Semant. Web **6**(2), 167–195 (2015). https://doi.org/10.3233/SW-140134

19. Li, D., Yang, S., Xu, K., Yi, M., He, Y., Wang, H.: Multi-task pre-training language model for semantic network completion. arXiv preprint arXiv:2201.04843 (2022). https://arxiv.org/abs/2201.04843

20. Liu, Z., Nguyen, T.K., Fang, Y.: Tail-GNN: tail-node graph neural networks. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 1109–1119. KDD 2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3447548.3467276

21. Ma, J., Guo, R., Wan, M., Yang, L., Zhang, A., Li, J.: Learning fair node representations with graph counterfactual fairness. In: Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pp. 695–703. WSDM 2022, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3488560.3498391

22. Malaviya, C., Bhagavatula, C., Bosselut, A., Choi, Y.: Commonsense knowledge base completion with structural and semantic context. Proc. AAAI Conf. Artif. Intell. **34**(03), 2925–2933 (2020). https://doi.org/10.1609/aaai.v34i03.5684

23. Miller, G.A.: WordNet: a lexical database for English. In: Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, 23–26 February 1992 (1992). https://aclanthology.org/H92-1116

24. Moon, S., Shah, P., Kumar, A., Subba, R.: OpenDialKG: explainable conversational reasoning with attention-based walks over knowledge graphs. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 845–854. Association for Computational Linguistics, Florence, Italy, July 2019. https://doi.org/10.18653/v1/P19-1081

25. Nan, G., Qiao, R., Xiao, Y., Liu, J., Leng, S., Zhang, H., Lu, W.: Interventional video grounding with dual contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2765–2775, June 2021

26. Nan, G., Zeng, J., Qiao, R., Guo, Z., Lu, W.: Uncovering main causalities for long-tailed information extraction. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 9683–9695. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, November 2021. https://doi.org/10.18653/v1/2021.emnlp-main.763

27. Nickel, M., Tresp, V., Kriegel, H.: A three-way model for collective learning on multi-relational data. In: Getoor, L., Scheffer, T. (eds.) Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28–July 2, 2011, pp. 809–816. Omnipress (2011). https://icml.cc/2011/papers/438_icmlpaper.pdf

28. Pearl, J., Glymour, M., Jewell, N.P.: Causal Inference in Statistics: A Primer. John Wiley & Sons, New York (2016)

29. Pearl, J., Mackenzie, D.: The Book of Why: The New Science of Cause and Effect. Basic Books (2018)

30. Qian, C., Feng, F., Wen, L., Ma, C., Xie, P.: 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), pp. 5434–5445. Association for Computational Linguistics, Online, August 2021. https://doi.org/10.18653/v1/2021.acl-long.422

31. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog (2019). https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

32. Radstok, W., Chekol, M.W., Schäfer, M.T.: Are knowledge graph embedding models biased, or is it the data that they are trained on? In: Kaffee, L., Razniewski, S., Hogan, A. (eds.) Proceedings of the 2nd Wikidata Workshop (Wikidata 2021) co-located with the 20th International Semantic Web Conference (ISWC 2021), Virtual Conference, October 24, 2021. CEUR Workshop Proceedings, vol. 2982. CEUR-WS.org (2021). https://ceur-ws.org/Vol-2982/paper-5.pdf

33. Shomer, H., Jin, W., Wang, W., Tang, J.: Toward degree bias in embedding-based knowledge graph completion. In: Proceedings of the ACM Web Conference 2023, pp. 705–715. WWW 2023, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3543507.3583544

34. Sun, Z., Deng, Z.H., Nie, J.Y., Tang, J.: Rotate: Knowledge graph embedding by relational rotation in complex space. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=HkgEQnRqYQ

35. Toutanova, K., Chen, D.: Observed versus latent features for knowledge base and text inference. In: Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality, pp. 57–66. Association for Computational Linguistics, Beijing, China, July 2015. https://doi.org/10.18653/v1/W15-4007

36. Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., Bouchard, G.: Complex embeddings for simple link prediction. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. In: Proceedings of Machine Learning Research, vol. 48, pp. 2071–2080. PMLR, New York, New York, USA, 20–22 June 2016. https://proceedings.mlr.press/v48/trouillon16.html

37. Wang, L., Zhao, W., Wei, Z., Liu, J.: SimKGC: simple contrastive knowledge graph completion with pre-trained language models. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4281–4294. Association for Computational Linguistics, Dublin, Ireland, May 2022. https://doi.org/10.18653/v1/2022.acl-long.295

38. Wang, X., Wang, D., Xu, C., He, X., Cao, Y., Chua, T.S.: Explainable reasoning over knowledge graphs for recommendation. Proc. AAAI Conf. Artif. Intell. **33**(01), 5329–5336 (2019). https://doi.org/10.1609/aaai.v33i01.33015329

39. Wang, X., et al.: KEPLER: a unified Model for Knowledge Embedding and Pretrained Language Representation. Trans. Assoc. Comput. Linguist. 9, 176–194 (2021). https://doi.org/10.1162/tacl_a_00360

40. Wang, Y., et al.: Should we rely on entity mentions for relation extraction? Debiasing relation extraction with counterfactual analysis. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3071–3081. Association for Computational Linguistics, Seattle, United States, July 2022. https://doi.org/10.18653/v1/2022.naacl-main.224

41. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: Brodley, C.E., Stone, P. (eds.) Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 27–31 July 2014, Québec City, Québec, Canada, pp. 1112–1119. AAAI Press (2014). http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8531

42. Xian, Y., Fu, Z., Muthukrishnan, S., de Melo, G., Zhang, Y.: Reinforcement knowledge graph reasoning for explainable recommendation. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 285–294. SIGIR 2019, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3331184.3331203

43. Yang, B., Yih, W., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015). http://arxiv.org/abs/1412.6575

44. Yang, Z., Liu, Y., Ouyang, C., Ren, L., Wen, W.: Counterfactual can be strong in medical question and answering. Inf. Process. Manage. **60**(4), 103408 (2023). https://doi.org/10.1016/j.ipm.2023.103408

45. Zamini, M., Reza, H., Rabiei, M.: A review of knowledge graph completion. Information. **13**(8), 396 (2022). https://doi.org/10.3390/info13080396

46. Zhang, D., Zhang, H., Tang, J., Hua, X.S., Sun, Q.: Causal intervention for weakly-supervised semantic segmentation. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 655–666. Curran Associates, Inc. (2020). https://proceedings.neurips.cc/paper_files/paper/2020/file/07211688a0869d995947a8fb11b215d6-Paper.pdf

47. Zhang, X., Cui, P., Xu, R., Zhou, L., He, Y., Shen, Z.: Deep stable learning for out-of-distribution generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5372–5382, June 2021

48. Zhang, Y., Dai, H., Kozareva, Z., Smola, A., Song, L.: Variational reasoning for question answering with knowledge graph. Proc. AAAI Conf. Artif. Intell. **32**(1), 1–8 (2018). https://doi.org/10.1609/aaai.v32i1.12057

49. Zhang, Z., Cai, J., Zhang, Y., Wang, J.: Learning hierarchy-aware knowledge graph embeddings for link prediction. Proc. AAAI Conf. Artif. Intell. **34**(03), 3065–3072 (2020). https://doi.org/10.1609/aaai.v34i03.5701

50. Zhao, T., Zhang, X., Wang, S.: GraphSmote: Imbalanced node classification on graphs with graph neural networks. In: Proceedings of the 14th ACM International Conference on Web Search and Data Mining. p. 833–841. WSDM 2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3437963.3441720

51. Zhao, T., Liu, G., Wang, D., Yu, W., Jiang, M.: Learning from counterfactual links for link prediction. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 26911–26926. PMLR, 17–23 July 2022. https://proceedings.mlr.press/v162/zhao22e.html

52. Zhu, Z., Zhang, Z., Xhonneux, L.P., Tang, J.: Neural bellman-ford networks: a general graph neural network framework for link prediction. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems, vol. 34, pp. 29476–29490. Curran Associates, Inc. (2021). https://proceedings.neurips.cc/paper_files/paper/2021/file/f6a673f09493afcd8b129a0bcf1cd5bc-Paper.pdf