TEF: Causality-Aware Taxonomy Expansion via Front-Door Criterion

Yuan Meng, Songlin Zhai, Yuxin Zhang, Zhongjian Hu, Guilin Qi[*]
School of Computer Science and Engineering, Southeast University, Jiangsu Province, China
Key Laboratory of New Generation Artificial Intelligence Technology and Its
Interdisciplinary Applications, Southeast University, Ministry of Education, China
`{yuan_meng, songlin_zhai, zzyx_cs, huzj, gqi}@seu.edu.cn`

Abstract

Taxonomy expansion is a primary method for enriching taxonomies, involves appending a large number of additional nodes (*i.e.,* queries) to an existing taxonomy (*i.e.,* seed), with the crucial step being the identification of the appropriate anchor (parent node) for each query by incorporating the structural information of the seed. Despite advancements, existing research still faces an inherent challenge of spurious query-anchor matching, often due to various interference factors (*e.g.,* the consistency of sibling nodes), resulting in biased identifications. To address the bias in taxonomy expansion caused by unobserved factors, we introduce the Structural Causal Model (SCM), known for its bias elimination capabilities, to prevent these factors from confounding the task through backdoor paths. Specifically, we employ the *Front-Door Criterion*, which guides the decomposition of the expansion process into a parser module and a connector. This enables the proposed causal-aware **T**axonomy **E**xpansion model to isolate confounding effects and reveal the true causal relationship between the query and the anchor. Extensive experiments on three benchmarks validate the effectiveness of TEF, with a notable 6.1% accuracy improvement over the state-of-the-art on the SemEval16-Environment dataset.

1 Introduction

Taxonomies systematically organize information into hierarchical structures (Wang et al., 2017a), providing semantic support for various downstream applications, *e.g.,* Personalized Recommendation (Wang et al., 2017b; Hu et al., 2021), Question Answering (Gupta et al., 2018; Yang et al., 2003) and Information Management (Nickerson et al., 2013; Sankepally, 2019). To accommodate the growth of knowledge, researchers focus on taxonomy expansion (Arous et al., 2023; Jiang et al., 2023; Cheng et al., 2022; Zhai et al., 2023), aiming to automatically enrich a seed taxonomy by appending a large number of queries without modifying its structure, as depicted in Figure 1. In this paper, we also follow this research line (Xia et al., 2023; Jiang et al., 2023; Xia et al., 2023; Zhai et al., 2023), seeking to advance the taxonomy expansion task via the structural causal model (SCM).

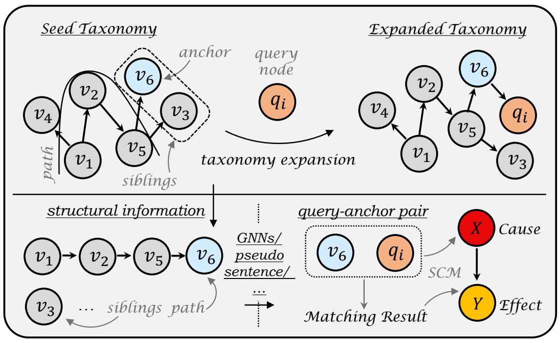


Figure 1: Illustration of the taxonomy expansion task from a causal perspective. The candidate query-anchor pair (e.g., (v_6, q_i)) and the corresponding matching result can be viewed as the *cause* and the *effect*, respectively. For instance, q_i ("banana") should ideally be attached under v_6 ("tropical fruits"), but due to its frequent co-occurrence with v_3 ("orange") as siblings and with v_5 ("fruit") as a parent-child relationship, it is more likely to be misclassified under v_5. This illustrates the interference from surrounding nodes.

The predominant trend in existing studies is to incorporate structural information from the seed taxonomy, such as egonets (Shen et al., 2020), mini-paths (Yu et al., 2020), the ancestral pseudo sentence (Jiang et al., 2022) or path prompts (Xia et al., 2023), thereby enhancing the matching performance between queries and anchors. However, these approaches still suffer from spurious query-anchor matching due to several interfering factors, leading to biased identifications. This ongoing challenge highlights the necessity for further research in this task.

---

[*] Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 8285–8294
January 19–24, 2025.

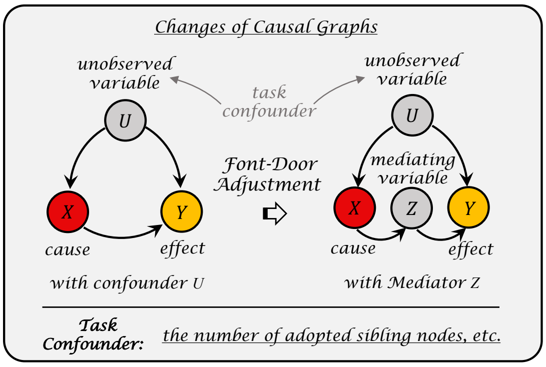


Figure 2: Alterations in causal graph structure under different scenarios. The left graph reflects the actual causal relation between the query-anchor pair (*i.e.,* cause X) and the matching result (*i.e.,* effect Y). To block the backdoor path ($X \leftarrow U \rightarrow Y$) , the *Front-Door Criterion* is employed, which introduces the mediating variable Z with its causal graph depicted on the right one.

Given the effectiveness of causal reasoning in bias analysis and the reliance on observational data in the taxonomy expansion task (Yuan et al., 2023; Zhang et al., 2023), we adopt the Structural Causal Model (SCM) Pearl (2009), which is well-known for its ability to mitigate bias from observational data through causal interventions. Following the insights of SCM, we construct a causal graph to deepen our understanding of the biased matching issue, where the query-anchor pair and its identification result represent the cause X and effect Y, respectively. As shown in Figure 2, the left causal graph reveals how prevailing approaches are vulnerable to confounding an unobserved variable U (*i.e.,* denoting interference factors) which opens the spurious backdoor path $X \leftarrow U \rightarrow Y$. This path introduces spurious statistical associations between X and Y, which underlie biased identifications (Pearl, 2019, 2009).

However, due to the elusive nature of the confounder U (Yuan et al., 2023; Zhang et al., 2023), directly blocking the backdoor path using backdoor adjustment from SCM is challenging (Pearl, 2019, 2009). Therefore, we propose a causal-aware **T**axonomy **E**xpansion model based on the ***Font-Door Criterion*** (Pearl, 2009; Pearl et al., 2016), termed TEF, where the mediating variable Z effectively block the backdoor path, as depicted in the right causal graph in Figure 2. Guided by this criterion, we decompose the whole expansion process into a mediator parser and a connector, effectively reducing the harmful effect of U on model predictions. Furthermore, we validate the effectiveness of the TEF model through experiments on three real-world benchmark datasets, demonstrating its robustness in matching state-of-the-art performance across most metrics. To summarize, the contributions of this paper are as follows:

- We introduce a causal perspective to mitigate query-anchor matching bias and present an innovative framework for potential future improvements.
- The proposed TEF model, grounded in the *Front-Door Criterion*, mitigates spurious query-anchor matching by blocking confounding backdoor paths with a mediating variable.
- Extensive experiments on three benchmark datasets demonstrate that TEF matches state-of-the-art performance across most metrics.

2 Preliminary

2.1 Task Definition

Taxonomy: A taxonomy, denoted as $\mathcal{T} = \{\mathcal{V}, \mathcal{E}\}$, is characterized as a directed acyclic graph (DAG) with the inherent hierarchical structure. It is comprised of a vertex set $\mathcal{V} = \{v_i | 1 \leq i \leq N\}$ and an edge set $\mathcal{E} = \{(v_i, v_j) \mid v_i \text{ is-a } v_j, 1 \leq i, j \leq N\}$ (Shen et al., 2020; Jiang et al., 2023; Zhai et al., 2023), where N denotes the total number of nodes in $\mathcal{V}$. For each pair of connected nodes $(v_i, v_j) \in \mathcal{E}$, an *is-a* relation (*e.g.,* v_i *is-a* v_j) is established, with v_j and v_i being the hypernym (parent) and hyponym (child), respectively. Specifically, the set of ancestral nodes in the taxonomy for a given node (*e.g.,* v_i) is denoted as $\mathcal{R}_i$.

Taxonomy Expansion: This task aims to append a large number of additional nodes (*i.e.,* queries) $\mathcal{Q} = \{q_j | 1 \leq j \leq M\}$ into an existing smaller seed taxonomy $\mathcal{T}^0$, thereby a larger taxonomy could be constructed. As illustrated in Figure 1, the critical step in this task is to perform the matching between a given query q_j and each node v_i in the seed $\mathcal{T}^0$ by determining the *is-a* relation between them (Wang et al., 2017a; Pannitto et al., 2017). As a formal definition, taxonomy expansion task aims to learn the probability distribution, formulated as:

$$P(Y = \textit{is-a} \,| X = (q_j, v_i)) \qquad (1)$$

where P indicates the learning distribution. X and Y are variables denoting the query-anchor pair and matching result, respectively.

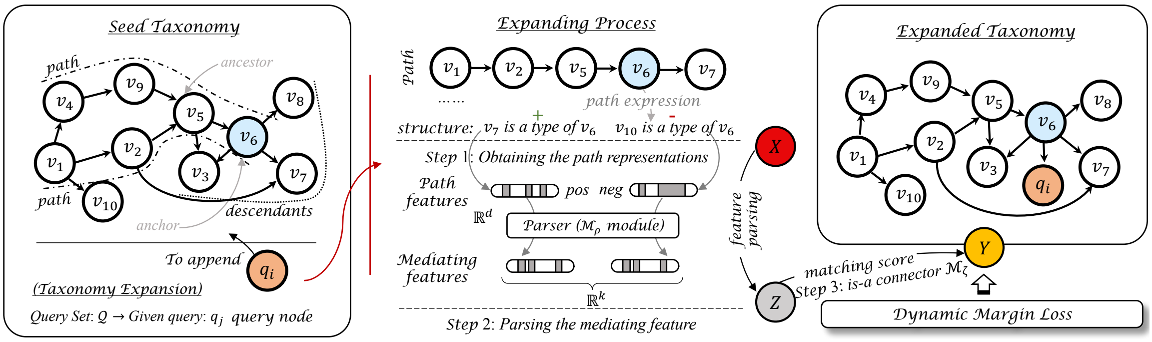


Figure 3: Architecture of the proposed TEF model. During TEF's expansion process, we strive to maintain its simplicity by only modeling essential paths (*e.g.,* $v_1 \to v_4 \to v_9 \to v_5 \to v_6$). To mitigate spurious query-anchor matching results, TEF employs the *front-Door Criterion* to block the backdoor path by the mediating variable with its feature parsed by a parser module (*i.e.,* $\mathcal{M}_\rho$).

2.2 Causal Analysis for Biased Identifications

As shown in Figure 2, the query-anchor pair (e.g., (q_j, v_i)) and its identification result (e.g., *is-a*) represent the cause X and effect Y, respectively. Formally, the task described in Eq. 1 can be defined as follows:

$$P(\textit{is-a}\ |(q_j, v_i)) := \mathcal{M}(f(\mathcal{T}^0, v_i, q_j)) \qquad (2)$$

where $\mathcal{M}$ denotes the learning model built on the seed taxonomy $\mathcal{T}^0$. f signifies the function that incorporates structural information from $\mathcal{T}^0$. As depicted in Figure 2, U serves as the confounder influencing the query features, which impacts the integration process f and leads to a suboptimal expansion model (e.g., $\mathcal{M}_1$).

To improve this task, extensive research employs superior f (*e.g.,* multi-view structural incorporation strategy (Jiang et al., 2022)) to achieve better results. However, directly addressing this problem is not the most effective approach, as the latent space of U is unobservable and unknown in advance (Yuan et al., 2023; Zhang et al., 2023), making traditional backdoor adjustment (Pearl, 2019; Pearl et al., 2016) infeasible in our task. To address this challenge, we adopt the *Front-Door Criterion*, training a neural network as the mediating variable Z to extract and transform latent is-a features from the original space into adjusted representations. For example, the model extracts features from "tropical fruit" to assess to assess whether it is a suitable hypernym for the query "banana". However, without proper adjustment, the model may mistakenly associate "fruit" with the query due to the shared features between "banana" and "orange", despite "banana" being a hyponym of "tropical fruit" rather than the specific match to the query.

3 Methodology

This section introduces the proposed TEF model, which leverages the *Front-Door Criterion* to mitigate identification bias. We begin by explaining the criterion's guidelines, followed by the TEF model's design. Lastly, we detail the model's components and describe the learning and inference process.

3.1 Guideline

Building on the Front-Door adjustment shown in Eq.**??**, the learning target introduced in Eq.1 can be reformulated as:

$$\begin{aligned} &P(Y = \textit{is-a}\ |X = (q_j, v_i)) \\ &\quad := P(Y = \textit{is-a}\ |do(X = (q_j, v_i))) \\ &\quad = \sum_z \{P(z|do(X = (q_j, v_i)))\cdot \\ &\qquad\qquad P(Y = \textit{is-a}\ |do(Z = z))\} \end{aligned} \qquad (3)$$

We can derive a guideline to reduce biased query-anchor detection by Eq. 3. Where the $do()$ simulates an intervention on a query-anchor pair (q_j, v_i), allowing the model to estimate the causal impact (i.e., the is-a relations) on the outcome Y.

Guideline. *It will be more advantageous to construct a model governed by the mediator Z, instead of directly developing a model to ascertain the relationship of query-anchor pairs since the original learning process is inherently susceptible to unobserved confounders.*

This guideline enlightens us to structure the model with two essential components, including a parser (designated as ρ-module) represented by $P(z|do(X = (q_j, v_i)))$, and the connector (referred to as the ζ-module) defined by $P(Y|do(Z = z))$. The motivation behind this guideline is

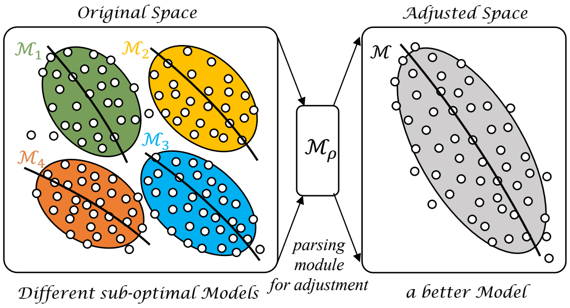


Figure 4: Illustration of the parsing process.

straightforward and powerful: ***we cast away all the unobserved factor (i.e., the confounder U) and focus solely on what can be controlled currently (i.e., the mediating variable Z).*** To summarize, TEF strives to learn two modules within the overall taxonomy expansion model $\mathcal{M}$:

$$\mathcal{M}(Y|X) := \begin{cases} \mathcal{M}_\rho := P(z|do(X = (q_j, v_i))) \\ \mathcal{M}_\zeta := P(Y|do(Z = z)) \end{cases} \tag{4}$$

where $\mathcal{M}_\rho$ and the $\mathcal{M}_\zeta$ are the two modules that need to be learned within the expansion model. As such, the working process of the model $\mathcal{M}$ could be delineated as:

$$\underbrace{X = (q_j, v_i) \rightarrow \mathcal{M}_\rho}_{\textit{parser}} \rightarrow \underbrace{\mathcal{M}_\zeta \rightarrow Y = \textit{is-a}}_{\textit{connector}}$$

In the following sections, we will expound on the implementations of these two modules.

3.2 ρ-Module: Parser

This module is central to the taxonomy expansion process, tasked with parsing the mediator features from the query-anchor representations. The primary objective of the parser is to map the query-anchor features into a unified, adjusted feature space, thereby minimizing the influence of the confounding factor U and alleviating biased identification. It can be defined as a matrix transformation of the original space (Gruenberg and Weir, 2013) , adjusted by a tailored parser function. This parser's formulation is shared across all taxonomy nodes, formulated as:

$$\mathcal{M}_\rho := \phi(\boldsymbol{W}_\rho \times g(q_j, v_i)^\top) + \boldsymbol{\epsilon} \tag{5}$$

where $\mathcal{M}_\rho$ is the mediator parser focusing on the adjustment of $\mathbb{R}^d \rightarrow \mathbb{R}^k$, where $\mathbb{R}^d$ and $\mathbb{R}^k$ correspond to the original d-dimensional and the adjusted k-dimensional feature spaces, respectively. $\boldsymbol{W}_\rho \in d \times k$ serves as the transfer matrix for adjusting the feature. $\boldsymbol{\epsilon} \in \mathbb{R}^k$ represents a Gaussian noised term (Guan et al., 2019) introduced to account for potential errors and variations in the parsing process. $\phi(\cdot)$ signifies the parsing function, and $g(q_j, v_i)$ is the feature extractor for the query-anchor pair based on the path containing v_i and q_j, facilitated by a pre-trained language model. This extractor returns the d-dimensional representations in the original space with $g(v_i, q_j)^\top$ being the transposition of generated features.

It is important to highlight that the parsing function $\phi(\cdot)$ is crucial to the whole parsing process, as it directly impacts the feasibility of the transition from $\mathbb{R}^d$ to $\mathbb{R}^k$. Following the causal identifiability in causal inference (Mooij et al., 2016; Shimizu et al., 2006), it can be observed that the non-linear parsing function could ensure the practicability of the parsing process given the Gaussian distributed feature space $\mathbb{R}^k$ (Hoyer et al., 2008). This deduction is succinct and ensures the robustness of the parser module $\mathcal{M}_\rho$. Equipped with this parser module, it is feasible to perform the expansion process independently from the observational datasets, thereby reducing the task bias. Regarding the function $g(v_i, q_j)$ in Eq. 5, it denotes a PLM, such as BERT, for generating the representation of the path containing the query q_j and the candidate node v_i. In order to align with the standard input format of PLMs, we adopt the method outlined in previous studies (Liu et al., 2021; Yu et al., 2020; Xu et al., 2022; Xia et al., 2023). This involves concatenating the nodes residing on the taxonomy path by the *is-a* relation, initiating from the root and ending with the last node. After encoding the textual path, a list of features can be obtained as follows:

$$[\boldsymbol{v}_{\text{[CLS]}}, \boldsymbol{q}_j, \boldsymbol{v}_1, ..., \boldsymbol{v}_{\text{root}}, \boldsymbol{v}_{\text{[SEP]}}]$$

where $\boldsymbol{v}_i, \boldsymbol{q}_j \in \mathbb{R}^d$ are the embedding vectors of node v_i and query q_j in the related taxonomy path. $\boldsymbol{v}_{\text{[CLS]}}$ symbolizes the representation of the classification token [CLS], utilized as the path feature.

An illustration of the parsing processing can be found in Figure 4, where the original and adjusted feature space indicate the results before and after the operation of the module $\mathcal{M}_\rho$, which introduces a nonlinear transformation to smoothly adjust the data distribution, thereby preventing excessive discretization and aiding in the construction of a more unified feature space. Specifically, the unadjusted approach might map training instances into distinct relationships, contingent upon the confounders, leading to sub-optimal variations of the expansion models, *e.g.,* $\mathcal{M}_1$ or $\mathcal{M}_2$ (as shown in Figure 4). By adjusting the feature space through the parser module, a unified representation is achieved, thereby mitigating the influence of confounders and enhancing the robustness of the resulting model, denoted as $\mathcal{M}$.

3.3 ζ-Module: Connector

In this module, the emphasis is placed on learning a connector between the adjusted features and the identification results. To maintain the model simplicity, we implement a scoring function designed to quantify the relevance of the query-anchor pair with respect to the *is-a* relation:

$$\mathcal{M}_\zeta := s(q_j, v_i) \tag{6}$$

where $\mathcal{M}_\zeta$ aims to perform the process: $\mathbb{R}^k \rightarrow \mathbb{R}^1$, where $\mathbb{R}^1$ indicates the 1-dimensional output space. $s(\cdot)$ is the scoring function acting the connector to bridge the parsed features and the *is-a* relevance of the query-anchor pair. $s(\cdot)$ is also simply implemented by a single Multi-Layer Perceptron (MLP) with the input dimension of k and the output dimension of 1.

3.4 Training and Inference

The primary objective of training the model $\mathcal{M}$ is to rank taxonomy nodes for a given query based on the score depicted in Eq. 6, thereby the optimal anchor could be ranked higher than other nodes. Formally, the optimization target aims to fulfill the following constraint for each query-anchor pair:

$$s(q_j, v_i^+) \geqslant s(q_j, v_l^-) + \gamma(q_j, v_i^+, v_l^-) \tag{7}$$

where $v_l^- \in \mathcal{V} - \mathcal{R}_i$. This equation serves as the cornerstone of our training methodology, where $s(q_j, v_i^+)$ and $s(q_j, v_l^-)$ denote the scores for the positive (v_i^+) and negative anchor (v_l^-) nodes, respectively. The margin function, $\gamma(q_j, v_i^+, v_l^-)$, enforces a gap between correct and incorrect scores to improve model discrimination. $\mathcal{R}_i$ denotes the ancestral set of v_i^+.

The margin function outlined in Eq. 7 encourages negative taxonomy-paths, which are relatively irrelevant to the query, to achieve lower scores. This is also evaluated by the Wu&P metric in next section. Consistent with Liu et al. (2021); Xia et al. (2023), we implement the margin function to ascertain the semantic similarity across diverse taxonomy paths, computed as:

$$r(q_j, v_i^+, v_l^-) = \left(\frac{|\mathcal{P}(q_j, v_i^+) \cup \mathcal{P}(q_j, v_l^-)|}{|\mathcal{P}(q_j, v_i^+) \cap \mathcal{P}(q_j, v_l^-)|} - 1\right) * \alpha \tag{8}$$

where α is a tunable parameter employed to adjust margins and set within the range of (0, 1]. $\mathcal{P}(q_j, v_i^+)$ and $\mathcal{P}(q_j, v_l^-)$ represent the positive and sampled negative paths, respectively. Specifically, for the positive path $\mathcal{P}(q_j, v_i^+)$, we select the corresponding taxonomy-path from the root node to the ground truth anchor v_i^+ with appending the given query q_j. Analogously, the negative one is constructed by appending the query q_j into a path that contains the randomly selected negative node $v_l^- \in \mathcal{V} - \mathcal{R}_i$.

Based on the learning target specified in Eq. 7, we could derive the loss function to measure the extent to which a non-parent node v_l^- violates the large-margin constraint of a query-anchor pair:

$$\begin{aligned}\mathcal{J}(\mathcal{T}) = \sum_{j=1}^{M}\sum_{l=1}^{L} \max\, [0, s(q_j, v_l^-) - s(q_j, v_i^+) \\ + \gamma(q_j, v_i^+, v_l^-)]\end{aligned} \tag{9}$$

where the function $\max\, [0, s(q_j, v_l^-) - s(q_j, v_i^+) + \gamma(q_j, v_i^+, v_l^-)]$ ensures that the large-margin constraint is fulfilled when $\mathcal{J} = 0$, indicating that there is no violation by negative samples. Otherwise, $\mathcal{J} > 0$ indicates a violation of the constraint.

During training, the mini-batch consists of pairs of samples, indicating that the positive and corresponding negative taxonomy-paths must be concurrently fed into the model. With these inputs, the margin function produces the corresponding margin; subsequently, we compute the margin loss and update the model parameters. For inference, the expansion model generates a scoring list of the nodes in seed taxonomy for a given query based on the function shown in Eq. 6. Armed with this ranking, we can obtain the predicted appropriate anchor and evaluate TEF's performance.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We evaluate TEF and the compared models on three publicly available English taxonomies from SemEval-2016 Task 13 (Bordea et al., 2016) in the domains of environment, science, and food. Table 1 details the statistics of these datasets. Following Liu et al. (2021); Yu et al. (2020); Xu et al. (2022); Xia et al. (2023), we construct training taxonomies in a top-down fashion, beginning from the root node and randomly expanding until we cover 80% of the nodes. These randomly-grown taxonomies are used as the seeds for self-supervised learning, with the remaining 20% leaf concepts allocated for testing.

4.1.2 Evaluation Metrics

During testing, the expansion models compute the matching score between a given query (*e.g.,* q_j) and all vertexes (*e.g.,* $v_i \in \mathcal{V}$) in the seed taxonomy. The nodes with the highest scores for each query are then selected to form a ranked list, denoted as $\{\hat{v}_1, \hat{v}_2, ..., \hat{v}_m\}$. This list can then be compared to the ground truth list $\{v_1, v_2, ..., v_m\}$. Following previous works (Liu et al., 2021; Yu et al., 2020; Xu et al., 2022; Xia et al., 2023), we adopt three standard metrics in taxonomy expansion for evaluation:

- **Accuracy** measures the exact match between predicted anchors and queries, formulated as:

  $$\text{Accuracy} = \frac{1}{M}\sum_{j=1}^{M}\mathbb{I}(\hat{v}_i = v_i)$$

  where $\mathbb{I}$ indicates the indicator function.

- **Mean Reciprocal Rank (MRR)** calculates the average of reciprocal ranks of the ground truth anchor for each query among all candidates, defined as:

  $$\text{MRR} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{\text{rank}(q_j)}$$

  where $\text{rank}(q_j)$ returns the position value of the ground truth anchor for query q_j.

- **Wu&Palmer Similarity (Wu&P)** calculates the semantic similarity between the predicted anchor $\hat{v}_j$ and the ground truth anchor v_j:

  $$\text{Wu\&P} = \frac{1}{M}\sum_{i=1}^{M}\frac{2 \times \#\text{depth}(\text{LCA}(\hat{v}_j, v_j))}{\#\text{depth}(\hat{v}_j) + \#\text{depth}(v_j)}$$

  where $\#\text{depth}(\cdot)$ indicates the depth of the input node and $\text{LCA}(\hat{v}_j, v_j)$ return the the last concept in the intersection of the paths from root to $\hat{v}_j$ and v_j.

Table 1: Statistics of three datasets. N and M denote the size of node and query sets with $|\mathcal{E}|$ indicating the number of edges. Additionally, #Depth and #$\overline{\text{Depth}}$ refer to the depth of taxonomies and the average depth of queries, respectively.

Dataset	N	M	$\lvert\mathcal{E}\rvert$	#Depth	#$\overline{\text{Depth}}$
Environment	261	201	261	6	3.78
Science	429	312	452	8	5.16
Food	1486	1184	1576	8	5.36

4.1.3 Compared Models

To fully assess the performance of the proposed TEF model, we conduct a comparative analysis against eight recent methods, including **BERT+MLP** (Xia et al., 2023), **HypeNet** (Shwartz et al., 2016), **TaxoExpan** (Shen et al., 2020), **ARBORIST** (Manzoor et al., 2020), **TMN** (Zhang et al., 2021), **TEMP** (Liu et al., 2021), **HEF** (Wang et al., 2021) and **TaxoPrompt** (Xu et al., 2022). These models range from methods that use pretrained BERT for feature extraction to those that incorporate more complex techniques, offering a comprehensive comparison of TEF's performance.

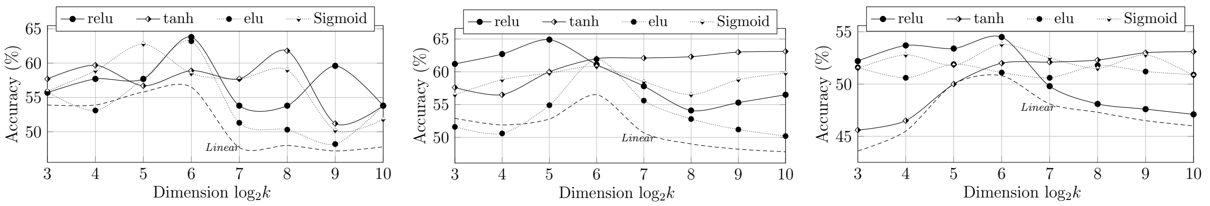
The experiments of all methods were conducted on an *NVIDIA A100-SXM4-40GB* machine, while the server is equipped with Intel(R) Xeon(R) Gold 5215 CPU @ 2.50GHz (80 cores) and 251 GB memory.

4.2 Performance Evaluation

Table 2 presents the empirical evaluation of the proposed TEF and all compared models across three different benchmark datasets using standard metrics in taxonomy expansion, including Accuracy, MRR and Wu&P similarity. Generally, the results gleaned from these experiments affirm the remarkable performance of TEF, as it outstripped the average-performing models on most evaluation metrics across all the datasets by a large margin. While the way TEF integrates the structural information is quite simple, the notable and consistent performance improvement it achieved is significant. This finding substantiates the hypothesis that the causal-aware TEF model significantly enhances taxonomy expansion and underscores the pivotal role of causal inference in diminishing spurious identifications and boosting the performance of taxonomy expansion task. Furthermore, despite a slight performance drop of the Wu&P metric on the SemEval16-Food is observed, it is important to

Datasets	SemEval16-Environment			SemEval16-Science			SemEval16-Food		
Metric	Accuracy(%)	MRR(%)	Wu&P(%)	Accuracy(%)	MRR(%)	Wu&P(%)	Accuracy(%)	MRR(%)	Wu&P(%)
BERT+MLP	11.1	21.5	47.9	11.5	15.7	43.6	10.5	14.9	47.0
TaxoExpan	11.1	32.3	54.8	27.8	44.8	57.6	27.6	40.5	54.2
HypeNet	16.7	23.7	55.8	15.4	22.6	50.7	20.5	27.3	63.2
STEAM	36.1	46.9	69.6	36.5	48.3	68.2	34.2	43.4	67.0
TMN	35.0	43.6	54.0	41.9	53.2	75.9	34.7	47.2	65.9
TEMP	49.0	62.0	75.9	54.4	64.6	84.6	45.2	57.1	78.3
HEF	55.3	65.3	71.4	53.6	62.7	75.6	47.9	55.5	73.5
TaxoPrompt	57.4	68.4	83.6	61.4	68.7	85.6	53.2	60.8	83.1
TEF	**63.5 *(+6.1)***	**72.8 *(+4.4)***	**86.1 *(+2.5)***	**64.3 *(+2.9)***	**73.4 *(+4.7)***	**88.1 *(+2.5)***	**54.1 *(+0.9)***	**62.5 *(+1.7)***	82.8 *(-0.3)*

Table 2: Experimental results of all models on three datasets for three evaluation metrics.



(a) SemEval16-Environment

(b) SemEval16-Science

(c) SemEval16-Food

Figure 5: Accuracy of TEF on three datasets with different k and parser functions ϕ.

highlight that there is still a considerable improvement in other evaluation metrics. Additionally, this drop could be potentially attributed to the simplicity of the method used to integrate the structural information in TEF. In summary, the promising results obtained from these experiments provide a solid foundation for future research endeavors that delve deeper into the integration of causal inference techniques within the realm of taxonomy expansion.

4.3 Effects of the Parser $\mathcal{M}_\rho$

In TEF, the parser module bridges the original suboptimal feature space and the adjusted space, enabling a more reliable quantification of the *is-a* relevance for query-anchor pairs. To ensure a fair comparison, we employ the same extractor as the baseline model, specifically Bert-base-uncased, for extracting the path feature vectors. Additionally, we conduct experiments to comprehensively assess the impact of this module, using TEF models in two variants: one incorporating the mediator and another without it, as summarized in Table 3. Notably, the TEF model incorporating $\mathcal{M}_\rho$ module consistently outperforms the model without it across all three SemEval16 datasets. Hence, the implementation of this module results in a significant improvement in the taxonomy expansion task, reflecting the effectiveness of it and advocating for its continued integration in this task. This positive impact of $\mathcal{M}_\rho$ on TEF's performance can be attributed to its role in effectively remitting the bias introduced by the unobserved confounders, thereby providing a nuanced way of capturing the intrinsic causal relationships between queries and their potential anchors and enhancing the model's ability to rank appropriate anchors accurately. To further investigate the details of performance improvements introduced by this module, we will experiment the parsing mechanism, including the parsing function ϕ and the dimension k of the parsing matrix $\boldsymbol{W}_\rho$.

4.4 Effects of ϕ and k

In this section, we further explore the impact of the parsing function ϕ and the dimension k on model performance and provide insights into the optimal choice of these parameters. The choice of ϕ has a significant effect on the capability of the parser model to capture better parsing results. To evaluate its influence, we experiment with different functions, each with different characteristics and keeping k constant. The results shown in Figure 5 demonstrate the model's sensitivity to ϕ. Generally, non-linear parsing functions achieve better results than linear analytic functions, showing the rationality of the non-linear functions. Furthermore, it's worth noting that certain non-linear functions demonstrate superior performance in terms of tax-

Table 3: The effects of mediator Z on three datasets.

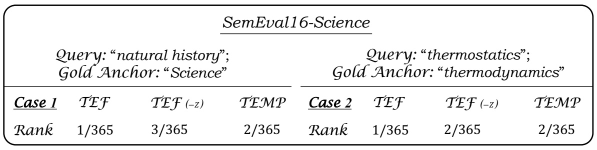
Accuracy(%)	TEF (-Z)	TEF	Ratio(↑%)
Environment	58.5	63.5	5.0
Science	59.4	64.3	4.9
Food	48.2	54.1	5.9

onomy expansion accuracy, whereas others may result in less favorable outcomes.

Meanwhile, the dimension of the adjusted feature space k is also an important factor that controls the granularity of feature representations. To investigate this, we set the value of k to 16, 32, and 64, while maintaining the analytical function ϕ constant. As illustrated in Figure 5, excessively low values of k result in performance degradation due to substantial information loss, whereas higher values of k preserve more information from the raw features, albeit at the expense of increased computational overhead. However, beyond a certain threshold, further increases in k yielded negligible improvements in model performance, suggesting a plateau or even a decline. Moreover, some parsing functions exhibited sensitivity to changes in dimensionality, while others remained stable.

4.5 Case Studies

Figure 6 further showcases the power of TEF by presenting the ranks assigned to two real queries within the SemEval16-Science dataset. Notably, TEF could rank the gold anchor higher than the baseline TEMP, showing its remarkable competency in reducing the bias in taxonomy expansion. Despite the varying familiarity of these two queries, the TEF model consistently predicts a high rank to the gold anchor, signifying its versatility and robustness. Comparing TEF with its variant TEF($-Z$), it becomes evident that the TEF's strength lies in its ability to effectively parse the mediator features for the purpose of adjusting the original feature space. This case study substantiates the intrinsic value and real-world applicability of our proposed TEF model in taxonomy expansion task, further strengthening the rationale for its broader adoption in related endeavors.



SemEval16-Science

Query: "natural history";
Gold Anchor: "Science"

Query: "thermostatics";
Gold Anchor: "thermodynamics"

Case 1	*TEF*	*TEF* (–z)	*TEMP*	**Case 2**	*TEF*	*TEF* (–z)	*TEMP*
Rank	1/365	3/365	2/365	*Rank*	1/365	2/365	2/365

Figure 6: Case study. TEF ($-Z$) indicates the model without the parser $\mathcal{M}_\rho$. The seed contains a total of 365 nodes.

5 Related Work

Taxonomy expansion (Zhai et al., 2023; Yu et al., 2020) has garnered significant attention for its ability to incorporate new nodes efficiently. Taxonomy expansion (Shen et al., 2020) introduces a self-supervised dynamic framework for effectively adding new concepts. Typically, existing approaches integrate the anchor's structural information, such as paths (Wang et al., 2021) or the hierarchical structure of taxonomies (Liu et al., 2021), to enhance the consistency of taxonomy expansion tasks. TaxoPrompt (Shen et al., 2020) uses a random walk to assimilate global structure via taxonomic prompts, enhancing taxonomy expansion performance. Additionally, as a powerful approach, TMN (Zhang et al., 2021) employs multiplicity scorers and a channel-wise gating mechanism to effectively tackle the one-to-pair matching problem. Similarly, TaxoEnrich (Jiang et al., 2022) is another advanced framework that effectively amalgamates semantic features and structural information from existing taxonomies, leading to enhanced performance in real-world datasets. While existing approaches incorporate seed taxonomy structure, they still struggle with spurious query-anchor matching, leading to biased outcomes. To address this, we propose TEF, a causality-oriented framework based on the Structural Causal Model (SCM) to identify genuine causal relationships, highlighting the potential of causal strategies for improving taxonomy expansion.

6 Conclusion

This paper pioneers the integration of causal inference into the taxonomy expansion task, introducing a novel perspective largely unexplored in previous research. We identified the challenge of spurious query-anchor matching and proposed the causal-aware TEF model to address this by effectively blocking the confounded backdoor path. Leveraging the *Front-Door Criterion*, we tackled the issue of unobserved confounders and implemented a crucial parser module within TEF. Despite its simplicity, experimental results on three real-world datasets demonstrated promising outcomes, highlighting the potential of incorporating causal inference in taxonomy expansion.

7 Limitation

Despite these contributions, we acknowledge the scope and limitation of our work. This paper is an initial attempt to explore the causal-aware approach for taxonomy expansion task. Our primary focus is to determine the applicability of the causal intervention theory in enhancing the reliability of expansion models. Therefore, in TEF, we strive to maintain simplicity by modeling only the essential paths and refraining from using complicated structural signals, such as different views of a graph (Jiang et al., 2022; Xia et al., 2023). While future researches could still delve into designing sophisticated techniques for better incorporating structural information, such investigations are beyond the purview of this paper. We hope our efforts could catalyze future inquires in this direction.

References

Ines Arous, Ljiljana Dolamic, and Philippe Cudré-Mauroux. 2023. Taxocomplete: Self-supervised taxonomy completion leveraging position-enhanced semantic matching. In *Proceedings of the ACM Web Conference, WWW*, pages 2509–2518, New York, NY, USA. ACM.

Georgeta Bordea, Els Lefever, and Paul Buitelaar. 2016. SemEval-2016 task 13: Taxonomy extraction evaluation (TExEval-2). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*, pages 1081–1091, San Diego, California. Association for Computational Linguistics.

Sijie Cheng, Zhouhong Gu, Bang Liu, Rui Xie, Wei Wu, and Yanghua Xiao. 2022. Learning what you need from what you did: Product taxonomy expansion with user behaviors supervision. In *the 38th IEEE International Conference on Data Engineering, ICDE*, pages 3280–3293, Kuala Lumpur, Malaysia. IEEE.

Karl W Gruenberg and Alan J Weir. 2013. *Linear geometry*, volume 49. Springer Science & Business Media, New York.

Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. 2019. Towards a deep and unified understanding of deep neural models in NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2454–2463, California, USA. PMLR.

Deepak Gupta, Rajkumar Pujari, Asif Ekbal, Pushpak Bhattacharyya, Anutosh Maitra, Tom Jain, and Shubhashis Sengupta. 2018. Can taxonomy help? improving semantic question matching using question taxonomy. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 499–513, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. 2008. Nonlinear causal discovery with additive noise models. In *NeurIPS*, pages 689–696. Curran Associates, Inc.

Xiaojiao Hu, Jiajie Xu, Weiqing Wang, Zhixu Li, and An Liu. 2021. A graph embedding based model for fine-grained poi recommendation. *Neurocomputing*, 428:376–384.

Minhao Jiang, Xiangchen Song, Jieyu Zhang, and Jiawei Han. 2022. Taxoenrich: Self-supervised taxonomy completion via structure-semantic representations. In *Proceedings of the ACM Web Conference*, WWW '22, page 925–934, New York, NY, USA. Association for Computing Machinery.

Song Jiang, Qiyue Yao, Qifan Wang, and Yizhou Sun. 2023. A single vector is not enough: Taxonomy expansion via box embeddings. In *Proceedings of the ACM Web Conference*, WWW '23, page 2467–2476, New York, NY, USA. Association for Computing Machinery.

Zichen Liu, Hongyuan Xu, Yanlong Wen, Ning Jiang, Haiying Wu, and Xiaojie Yuan. 2021. TEMP: taxonomy expansion with dynamic margin loss through taxonomy-paths. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 3854–3863, Virtual Event / Punta Cana, Dominican Republic. Association for Computational Linguistics.

Emaad A. Manzoor, Rui Li, Dhananjay Shrouty, and Jure Leskovec. 2020. Expanding taxonomies with implicit edge semantics. In *WWW: The Web Conference 2020*, pages 2044–2054, Taipei, Taiwan. ACM / IW3C2.

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: Methods and benchmarks. *J. Mach. Learn. Res.*, 17(1):1103–1204.

Robert C Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22(3):336–359.

Ludovica Pannitto, Lavinia Salicchi, and Alessandro Lenci. 2017. Ahyda: Automatic hypernym detection with feature augmentation. In *Italian Conference on Computational Linguistics*, pages 11–12, Rome. Torino: Accademia University Press.

J. Pearl, M. Glymour, and N.P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. Wiley, University of California Los Angeles, USA.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press, New York.

Judea Pearl. 2019. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3):54–60.

Rashmi Sankepally. 2019. Event information retrieval from text. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1447, New York, NY, USA. Association for Computing Machinery.

Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. 2020. Taxoexpan: Self-supervised taxonomy expansion with position-enhanced graph neural network. In *WWW'20*, pages 486–497, Taipei, Taiwan. ACM / IW3C2.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. 2006. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 2389–2398, Berlin, Germany. The Association for Computer Linguistics.

Chengyu Wang, Xiaofeng He, and Aoying Zhou. 2017a. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In *the Conference on Empirical Methods in Natural Language Processing*, pages 1190–1203, Copenhagen, Denmark. Association for Computational Linguistics.

Suyuchen Wang, Ruihui Zhao, Xi Chen, Yefeng Zheng, and Bang Liu. 2021. Enquire one's parent and child before decision: Fully exploit hierarchical structure for self-supervised taxonomy expansion. In *WWW'21*, pages 3291–3304, Virtual Event / Ljubljana, Slovenia. ACM / IW3C2.

Weiqing Wang, Hongzhi Yin, Ling Chen, Yizhou Sun, Shazia Sadiq, and Xiaofang Zhou. 2017b. St-sage: A spatial-temporal sparse additive generative model for spatial item recommendation. *ACM Transactions on Intelligent Systems and Technology*, 8(3):1–25.

Fei Xia, Yixuan Weng, Shizhu He, Kang Liu, and Jun Zhao. 2023. Find parent then label children: A two-stage taxonomy completion method with pre-trained language model. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023*, pages 1032–1042, Dubrovnik, Croatia. Association for Computational Linguistics.

Hongyuan Xu, Yunong Chen, Zichen Liu, Yanlong Wen, and Xiaojie Yuan. 2022. Taxoprompt: A prompt-based generation method with taxonomic context for self-supervised taxonomy expansion. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4432–4438, Vienna, Austria. IJCAI.

Hui Yang, Tat-Seng Chua, Shuguang Wang, and Chun-Keat Koh. 2003. Structured use of external knowledge for event-based open domain question answering. In *ACM SIGIR*, SIGIR '03, page 33–40, New York, NY, USA. Association for Computing Machinery.

Yue Yu, Yinghao Li, Jiaming Shen, Hao Feng, Jimeng Sun, and Chao Zhang. 2020. STEAM: self-supervised taxonomy expansion with mini-paths. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1026–1035, Virtual Event, CA, USA,. ACM.

Siyu Yuan, Deqing Yang, Jinxi Liu, Shuyu Tian, Jiaqing Liang, Yanghua Xiao, and Rui Xie. 2023. Causality-aware concept extraction based on knowledge-guided prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 9255–9272, 1 Harbour Square, Toronto, Canada. Association for Computational Linguistics.

Songlin Zhai, Weiqing Wang, Yuanfang Li, and Yuan Meng. 2023. Dng: Taxonomy expansion by exploring the intrinsic directed structure on non-gaussian space. *Proceedings of the AAAI Conference*, 37(5):6593–6601.

Jieyu Zhang, Xiangchen Song, Ying Zeng, Jiaze Chen, Jiaming Shen, Yuning Mao, and Lei Li. 2021. Taxonomy completion via triplet matching network. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI*, pages 4662–4670, Virtual Event. AAAI Press.

Zaixi Zhang, Qi Liu, Zhicai Wang, Zepu Lu, and Qingyong Hu. 2023. Backdoor defense via deconfounded representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12228–12238, Vancouver Canada. IEEE.