TaxoPrompt: A Prompt-based Generation Method with Taxonomic Context for Self-Supervised Taxonomy Expansion

Hongyuan Xu[1,2], **Yunong Chen**[1,2], **Zichen Liu**[1,2], **Yanlong Wen**[1,2*] and **Xiaojie Yuan**[1,2]
[1]College of Computer Science, Nankai University
[2]Tianjin Media Computing Center, Nankai University
{xuhongyuan, chenyunong, liuzichen}@dbis.nankai.edu.cn, {wenyl, yuanxj}@nankai.edu.cn

Abstract

Taxonomies are hierarchical classifications widely exploited to facilitate downstream natural language processing tasks. The taxonomy expansion task aims to incorporate emergent concepts into the existing taxonomies. Prior works focus on modeling the local substructure of taxonomies but neglect the global structure. In this paper, we propose TaxoPrompt, a framework that learns the global structure by prompt tuning with taxonomic context. Prompt tuning leverages a template to formulate downstream tasks into masked language model form for better distributed semantic knowledge use. To further infuse global structure knowledge into language models, we enhance the prompt template by exploiting the taxonomic context constructed by a variant of the random walk algorithm. Experiments on seven public benchmarks show that our proposed TaxoPrompt is effective and efficient in automatically expanding taxonomies and achieves state-of-the-art performance.

1 Introduction

Taxonomy, a tree structure of hierarchical classifications for a given set of objects, is widely used in several NLP downstream tasks such as query understanding [Yang *et al.*, 2017], information extraction [Karamanolakis *et al.*, 2020], and personalized recommendation [Huang *et al.*, 2019]. However, the low coverage problem remains a bottleneck that restricts the performance of these taxonomy-dependent applications. Recent studies [Shen *et al.*, 2018; Wang *et al.*, 2021] focus on the automatic taxonomy expansion task to cover emergent concepts since manually curating a taxonomy is labor-intensive, domain-specific, and time-consuming. The taxonomy expansion task aims to insert new concepts ("query concepts") into an existing taxonomy ("seed taxonomy") by finding their most appropriate hypernyms ("anchor concepts" or "positions") in the seed taxonomy while maintaining the consistency of the expanded taxonomy.

Early taxonomy expansion methods focus on learning semantic and contextual features of query concepts and anchor nodes extracted from corpora [Nickel and Kiela, 2017; Shen *et al.*, 2018]. However, these methods only model the hypernym-hyponym (*is-a*) relations but fail to capture the structure information of the existing taxonomy. To better leverage the existing taxonomy, recent works model the designed heuristic local structures of taxonomies which contain richer hierarchical information [Shen *et al.*, 2020; Yu *et al.*, 2020; Wang *et al.*, 2021]. Nevertheless, they all neglect the global structure information of the existing taxonomy and only consider the relations between query concepts and anchor structures.

*Corresponding author.

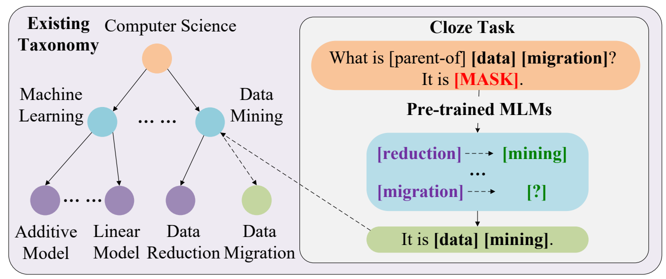

Figure 1: An example of the prompt-based hypernym generation. The grey box (right) shows the example of attaching the new concept "Data Migration" to the existing "Computer Science" taxonomy.

To overcome the above limitations, we propose TaxoPrompt, a prompt-based taxonomy expansion framework. Proven to be effective for capturing the global structure information of graphs [Chen *et al.*, 2020], the random walk is leveraged for our framework to generate self-supervision signals. Specifically, based on the characteristics of taxonomies, we design a random walk algorithm with different walk types. The walked paths generated from the existing taxonomy construct taxonomic context as our self-supervision signals.

Inspired by recent successes of prompt-based methods [Liu *et al.*, 2021a], we employ the prompt tuning paradigm to fully exploit the semantic knowledge in the language model (LM). Under the prompt paradigm, we formulate the taxonomy expansion problem as a hypernym generation task. As shown in Figure 1, TaxoPrompt applies the prompt template designed for hypernym generation to enhance the learning of lexical-syntactic features. To make the best use of constructed self-supervision signals, we complement the prompt template by attaching taxonomic context as knowledgeable context during training. In this way, we infuse the global structure knowledge into the language model. TaxoPrompt tends to generate hypernyms with the structure consistency after learning the structure knowledge of the existing taxonomy.

Our contributions are summarized as follows:

- We propose a self-supervised framework that expands taxonomy by prompt-based hypernym generation. The framework reduces the time complexity that previously increased with the square of the number of nodes to linear in both training and inferring.
- We design a random walk algorithm to capture the global structure of the existing taxonomy and infuse structure knowledge into the LM in a contextual way.
- Extensive experiments on seven benchmark taxonomy datasets demonstrate the efficiency and effectiveness of our method.

2 Related Work

Taxonomy Expansion. The taxonomy expansion methods aim to attach emergent concepts to the most appropriate anchor node in seed taxonomies. Many recent methods achieved considerable success. For example, TaxoExpan [Shen *et al.*, 2020] proposed position-enhanced egonet for neighborhood information aggregation and HyperExpan [Ma *et al.*, 2021] further extended such approach to hyperbolic space. STEAM [Yu *et al.*, 2020] serialized the existing taxonomies into mini-paths and scored the query node with them. TEMP [Liu *et al.*, 2021b] exploited the taxonomy-path to model hierarchical information. HEF [Wang *et al.*, 2021] designed a novel ego-tree structure to exploit hierarchical structure fully. To sum up, structure information is important for taxonomy expansion. Our method models the global structure of seed taxonomy and infuses global structure knowledge into the LM for better expansion.

Different Scenarios. Some recent methods focused on expansion tasks in different scenarios. Arborist [Manzoor *et al.*, 2020] first studied heterogeneous semantics in taxonomies. TMN [Zhang *et al.*, 2021] proposed a taxonomy completion task where new concepts can be placed between existing nodes. GenTaxo [Zeng *et al.*, 2021] enhanced the taxonomy completion by generating appropriate concept names to complement taxonomies. TaxoOrder [Song *et al.*, 2021] researched the importance of discovering hypernym-hyponym relations among new concepts before attaching them. Musubu [Takeoka *et al.*, 2021] addressed the low-resource problem using LMs. In this paper, we focus on the prompting solution for the leaf expansion task.

Tuning Paradigm. Pre-trained LMs have been widely exploited in taxonomy expansion task [Yu *et al.*, 2020; Takeoka *et al.*, 2021; Liu *et al.*, 2021b; Wang *et al.*, 2021]. Most existing methods followed a fine-tuning paradigm where LM is adapted to the downstream tasks like binary classification. Such a paradigm is prone to catastrophic forgetting, where the LM may lose its acquired knowledge before fine-tuning [Liu *et al.*, 2021a]. In our work, we follow a prompt tuning paradigm and adapt the taxonomy expansion task to LMs for better knowledge utilization.

3 Methodology

3.1 Preliminary

Definition 1 (Taxonomy)**.** *We follow [Zhang* et al.*, 2021] and define a taxonomy $\mathcal{T} = (\mathcal{N}, \mathcal{E})$ as a directed acyclic graph where each node $n \in \mathcal{N}$ represents a concept (i.e., a word or a phrase) and each directed edge $\langle u, v\rangle \in \mathcal{E}$ implies a general "is a hyponym of" relation or heterogeneous relations such as "is type of" or "is capital of". The taxonomy follows a hierarchical structure where concept u is the most specific concept related to the concept v. Note that a concept node may have multiple parents in a large-scale taxonomy.*

Definition 2 (Taxonomy Expansion)**.** *Given (1) an existing taxonomy $\mathcal{T}^0 = \left(\mathcal{N}^0, \mathcal{E}^0\right)$ and (2) a set of new concepts $\mathcal{C}$, which can be either manually specified or automatically extracted from corpus $\mathcal{D}$. The main goal of taxonomy expansion task is to complete the existing taxonomy $\mathcal{T}^0$ into a larger taxonomy $\mathcal{T} = \left(\mathcal{N}^0 \cup \mathcal{C}, \mathcal{E}^0 \cup \mathcal{R}\right)$ with $\mathcal{R}$ being the newly discovered relations for each concept $c \in \mathcal{C}$.*

In this paper, we solve the taxonomy expansion task by generating hypernym for a query concept. More specifically, given (1) a set of terms $\mathcal{N}^0$ and (2) a new concept $c \in \mathcal{C}$, our goal is to generate a list of tokens $\mathcal{L} = \left(\ell_1, \ell_2, \ldots, \ell_{|\mathcal{L}|}\right)$, where $\ell_i \in \mathcal{V}$ is i-th token and $|\mathcal{L}|$ is the total length of the generated list, $\mathcal{V}$ denotes the token vocabulary. Then, we convert the token list $\mathcal{L}$ to a concept $u \in \mathcal{N}^0$ and add $\langle u, c\rangle$ to the existing taxonomy. Mathematically, our final taxonomy expansion goal can be formulated as following $|\mathcal{C}|$ independent optimization problems [Shen *et al.*, 2020]:

$$\hat{u_i} = \underset{u_i \in \mathcal{N}^0}{\arg\max} \log \mathrm{P}\left(c_i \mid u_i, \Theta\right), \forall i \in \{1, 2, \ldots, |C|\}, \quad (1)$$

where Θ is the set of model parameters and u_i is the hypernym generated for the query concept c_i.

3.2 Modeling Hypernym Generation

Backbone Generation Model

TaxoPrompt follows the prompt tuning paradigm [Liu *et al.*, 2021a] and exploits LMs in the masked language model task way (shown in Figure.1). Specifically, TaxoPrompt takes BERT_{base} [Devlin *et al.*, 2019] as its inner LM. The impact of choices for LMs will be discussed in Section 4.4.

TaxoPrompt first leverages a prompting function [Schick and Schütze, 2021] to modify an input query concept c into a base prompt $\mathcal{P}(c)$. As shown in Figure 2, the function applies a template with two slots: "What is parent-of [X]? It is [MASK]." and fills slot [X] with the name of input concept c:

$$\mathcal{P}(c) = \text{What is parent-of c? It's [MASK].}$$

Then, TaxoPrompt feeds the prompt into the LM for word tokenization using algorithm like WordPiece [Schuster and Nakajima, 2012] and gets a prompt sentence s with n tokens:

$$s = t_1, t_2, \ldots, t_i, \langle\text{mask}\rangle_1, \ldots, \langle\text{mask}\rangle_{|\mathcal{L}|}, t_j, \ldots, t_n, \quad (2)$$

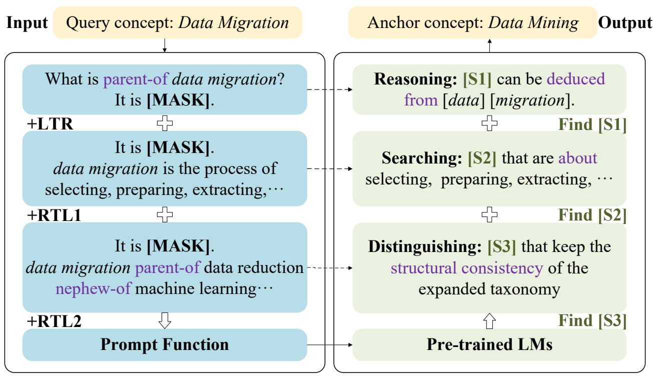

Figure 2: A pipeline of the hypernym generation. The left box shows the construction process of the prompt sentences (note that **LTR** is short for **L**eft-**T**o-**R**ight context, and **RTL** is the opposite.), and the right box illustrates the function of LMs (the dashed arrows indicate the correspondence, **[S]** represents the possible answer set).

where $t_{1:i}$ is the left-to-right context for masked positions and $t_{j:n}$ is the opposite. Finally, the LM is applied for parallel masked positions prediction by calculating the conditional probability distribution [Jiang *et al.*, 2020]:

$$\hat{t_k} = \underset{t'_k \in \mathcal{V}}{\operatorname{argmax}} P\left(t'_k \mid t_1, \ldots, t_{k-1}, t_{k+1}, \ldots, t_n, \Theta\right), \quad (3)$$

where t'_k indicates the generated token for k-th position in Eq.(2), and $t_{1:k-1}, t_{k+1:n}$ are surrounding tokens which can either be words or mask tokens. $P(\cdot, \Theta)$ can be measured with logit scores :

$$\begin{cases} O = \text{LayerNorm}(\sigma(HD^T)) \\ L = OM^T \end{cases}, \quad (4)$$

where $H \in \mathbb{R}^{n \times h}$ is the output of last multi-head self-attention layer and h represents the hidden size. $D \in \mathbb{R}^{h \times h}$ and $M \in \mathbb{R}^{|\mathcal{V}| \times h}$ are learnable projection matrices. σ represents the activation function. $L \in \mathbb{R}^{n \times |\mathcal{V}|}$ equals to the logits scores, and we denote $L(k,t)$ as the logit score of token t at k-th position of the prompt sentence s. Thus, Eq.(3) is equivalent to:

$$\hat{t_k} = \underset{t'_k \in \mathcal{V}}{\operatorname{argmax}} L(k, t'_k), \quad (5)$$

after $|\mathcal{L}|$ times prediction, the token list $\mathcal{L}$ is generated.

Discussion. We believe that our prompt-based hypernym generation method is sufficient to perform lexical-syntactic reasoning for two reasons: (1) Lexical-syntactic features [Yu *et al.*, 2020] are shown to LMs through tokenization algorithm [Liu *et al.*, 2021b]; (2) LMs have acquired semantic meanings and contextual relations of tokens after pre-trained on a large corpus [Takeoka *et al.*, 2021]. Prompt learning can make better use of existing knowledge.

3.3 Knowledgeable Context Construction

We note that no right-to-left context is available for masked positions in Eq.(2), which underutilizes the powerful bidirectional MLM task. Besides, task-specific knowledge like structure knowledge is also very important to the expansion task. In this section, we introduce an approach to infuse knowledge into the LM by attaching knowledgeable contexts to the base prompt.

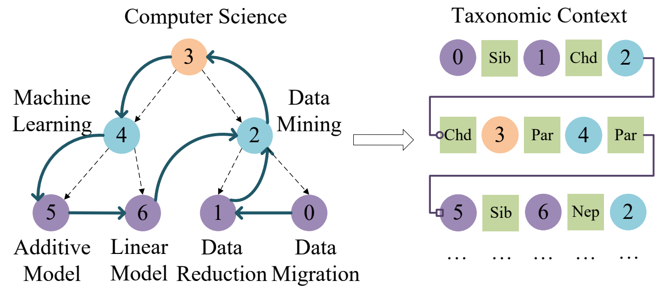

Figure 3: An illustration of the proposed random walk algorithm for one taxonomic context construction. Note that circles represent concept nodes, and squares represent walks of different relations. The dash lines indicate hypernym-hyponym relations.

Taxonomic Context. Previous approaches focus on modeling the local substructure of taxonomies like ego-nets[Shen *et al.*, 2020], mini-paths[Yu *et al.*, 2020] and ego-trees[Wang *et al.*, 2021]. In contrast, we model the global structure of taxonomies by the following steps: First, we define a list of relation tokens $\mathcal{R}$, where $\mathcal{R} = \{$**parent-of, child-of, sibling-of, nephew-of, posterity-of**$\}$. Then, we design a three-stage random walk algorithm: (1) Selecting a random relation r_i from $\mathcal{R}$, where i indicates the i-th walk. (2) Choosing a concept u_i randomly from the set consisting of concepts that hold relation r_i with u_{i-1}, where u_{i-1} is the last concept in current path. (3) Attaching r_i and u_i to the tail of current path. After κ times random walk, we serialize the taxonomy into a walked path named taxonomic context (shown in Figure 3):

$$\mathcal{W}(c) = c, r_1, u_1, r_2, \ldots, u_{i-1}, r_i, u_i \ldots, u_{\kappa-1}, r_\kappa, u_\kappa, \quad (6)$$

where $\mathcal{W}(c)$ denotes one taxonomic context for a query concept $c \in \mathcal{N}^0$ and κ is a hyperparameter. Finally, we concatenate τ taxonomic contexts to construct the full taxonomic context $\mathcal{W}'(c)$. Suppose the query concept is "*Data Migration*", a possible taxonomic context could be "*Data Migration **parent-of** Data Reduction **nephew-of** Machine Learning*" with $\kappa = 2, \tau = 1$. After seeing all taxonomic contexts, the LM learns the global structure knowledge of taxonomy by capturing hierarchical information and understanding relations between local structures.

Descriptive Context. Corpora resources like Wikipedia summary [Liu *et al.*, 2021b] and WordNet definition [Wang *et al.*, 2021] have been proved to imply the target *is-a* relation. We denote the description of a query concept c as $\mathcal{D}(c)$. [He *et al.*, 2020] has demonstrated that the LM can be complemented by $\mathcal{D}(c)$ since the LM learns to summarize the main attributes from the description:

$$\mathcal{D}(c) \xrightarrow{LM} \{a_1, a_2, \ldots, a_n\}$$

where a_i is one summarized attribute of concept c. In this way, the LM builds a deeper understanding of concept semantics. Finally, our prompt function of the query node c

during training is formulated as:

$$\mathcal{P}\left(c\right)[SEP]\mathcal{D}\left(c\right), \\ \mathcal{P}\left(c\right)[SEP]\mathcal{W}'\left(c\right) \tag{7}$$

and the former prompt is applied during inference.

3.4 Learning and Inference

Training Data Construction. Given one edge $\langle u, c\rangle$ from the existing taxonomy $\mathcal{T}^0 = \left(\mathcal{N}^0, \mathcal{E}^0\right)$, we first construct prompt for query concept c using the prompt function in Eq.(7). Then we generate the prompt sentence s in Eq.(2) by feeding the prompt to the LM tokenizer. Answer token list $\mathcal{L}_{gold}$ is constructed by tokenizing the parent node u. Notice that the number of masked tokens in s equals to $|\mathcal{L}_{gold}|$. Finally, one training instance $X = \langle s, \mathcal{L}_{gold}\rangle$ corresponds to the edge $\langle u, c\rangle$ is created. By repeating the above process for each edge in $\mathcal{T}^0$, we obtain the full training dataset $\mathbb{X} = \{X_1, X_2, \ldots, X_{|\mathcal{E}^0|}\}$.

Model Training. We adopt cross entropy loss as the main training objective:

$$\mathcal{L}(\Theta) = -\frac{1}{|\mathbb{X}|}\sum_{X_i\in\mathbb{X}}\left[\sum_{t_k^*\in\mathcal{L}_{gold}}\log\frac{\exp\left(L\left(k, t_k^*\right)\right)}{\sum\limits_{t\in\mathcal{V}}\exp\left(L\left(k,t\right)\right)}\right], \tag{8}$$

where t_k^* represents the ground truth token at the k-th position of s and L is logit scores defined in Eq.(4). The above equation is also known as MLM loss.

Inference. During inference, for each new concept $c \in \mathcal{C}$ and an candidate concept $u \in \mathcal{N}_0$, we construct a prompt sentence s without taxonomic context and calculate their match score by the average logit score:

$$\text{score}\left(u, c\right) = \frac{1}{|\mathcal{L}_u|}\sum_{\ell_i\in\mathcal{L}_u} L\left(k, \ell_i\right), \tag{9}$$

where $\mathcal{L}_u = \text{tokenize}\left(u\right)$ and ℓ_i represents i-th token in $\mathcal{L}_u$. k-th position of s is the i-th mask token, where ℓ_i is supposed to be filled in.

Complexity Analysis. The time complexity of training is $\mathcal{O}(I \cdot |\mathcal{E}^0| \cdot l_{avg}^2 \cdot d)$, where I is the number of iterations, l_{avg} is the average length of input sentence for the LM and d is the dimension of embedding. We infuse global structure knowledge into the LM to distinguish similar positions instead of negative sampling, making it possible to train efficiently. The time complexity of inference is $\mathcal{O}(|C| \cdot l_{avg}^2 \cdot d)$, where $|C|$ is the total number of new concepts, while the time complexity of the previous transformer-based methods[Liu *et al.*, 2021b; Wang *et al.*, 2021] is $\mathcal{O}(|C|^2 \cdot l_{avg}^2 \cdot d)$.

4 Experiments

In this section, we evaluate the performance of our proposed method TaxoPrompt. Our experiments are designed to answer the following research questions (RQs):

- **RQ1:** How does TaxoPrompt model perform compared with state-of-the-art taxonomy expansion methods?
- **RQ2:** How do different components (i.e., base prompt, descriptive context, and taxonomic context) affect TaxoPrompt?
- **RQ3:** What is the impact of different prompt designs (i.e., choice of language models and design of prompt template)?

Dataset	$\lvert\mathcal{N}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{D}\rvert$
Environment	261	261	6
Science	429	452	8
Food	1,486	1,576	8
MAG-CS	24,754	42,329	6
MAG-PSY	23,187	30,041	6
WordNet-Verb	13,936	13,408	13
WordNet-Noun	83,073	76,812	20

Table 1: Dataset Statistics. $|\mathcal{N}|$ and $|\mathcal{E}|$ are the number of nodes and edges in the existing taxonomy. $|\mathcal{D}|$ indicates the taxonomy depth.

4.1 Experimental Setups

Datasets

We evaluate our model on different benchmarks. The statistics of each dataset are shown in Table 1.

Low-resource Taxonomies. Following previous work [Yu *et al.*, 2020; Liu *et al.*, 2021b; Wang *et al.*, 2021], we evaluate our TaxoPrompt on three benchmark taxonomies from SemEval-2016 Task 13[Bordea *et al.*, 2016]. We experiment on three English datasets from different domains: environment, science, and food. We follow [Wang *et al.*, 2021] and exclude 20% nodes in each dataset, of which ten nodes are separated as the validation set and the rest as the test set.

Large-scale Taxonomies. Following previous work[Shen *et al.*, 2020; Zhang *et al.*, 2021; Ma *et al.*, 2021], we further evaluate our model on four large-scale real-world taxonomies from Microsoft Academic Graph (MAG) [Sinha *et al.*, 2015] and WordNet [Jurgens and Pilehvar, 2016]. We randomly sample 1,000 leaf nodes for each dataset as the test set and another 1,000 leaves as the validation set.

Evaluation Metrics

TaxoPrompt ranks all candidate hypernyms by calculating the score in Eq.(9) during testing. Given n query nodes, we denote their ground truth hypernyms as $\{\hat{u_1}, \hat{u_2}, \ldots, \hat{u_n}\}$ and the predicted hypernyms as $\{u_1, u_2, \ldots, u_n\}$ for low-resource benchmarks. Following prior works[Yu *et al.*, 2020; Shen *et al.*, 2020; Liu *et al.*, 2021b; Wang *et al.*, 2021], we adopt the following metrics:

(1) **Accuracy (Acc)** measures the times when the predicted hypernym exactly equals to the ground truth:

$$\text{Acc} = \frac{1}{n}\sum_{i=1}^{n}\left(u_i = \hat{u}_i\right)$$

(2) **Mean reciprocal rank (MRR)** calculates the average of reciprocal ranks with:

$$\text{MRR} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\text{rank}\left(\hat{u}_i\right)}$$

Dataset	Environment			Science			Food		
Metric	Acc	MRR	Wu&P	Acc	MRR	Wu&P	Acc	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
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	49.2	63.5	<u>77.7</u>	<u>57.8</u>	<u>67.5</u>	<u>85.3</u>	47.6	<u>60.5</u>	<u>81.0</u>
HEF	<u>55.3</u>	<u>65.3</u>	71.4	53.6	62.7	75.6	<u>47.9</u>	55.5	73.5
TaxoPrompt	**57.4**	**68.4**	**83.6**	**61.4**	**68.7**	**85.6**	**53.2**	**60.8**	**83.1**

Table 2: Overall experimental results on low-resource datasets (in %). We report our performance using the average of three runs. Note that we highlight the best results and underline the second best.

(3) **Wu & Palmer similarity (Wu&P)** represents the semantic similarity between the prediction and the ground truth:

$$\text{Wu\&P} = \frac{1}{n}\sum_{i=1}^{n}\frac{2\,|u_i \cap \hat{u}_i|}{|u_i| + |\hat{u}_i|},$$

where $|\cdot|$ indicates the depth of a concept in the taxonomy and $u_i \cap \hat{u}_i$ is the last concept in the intersection of the paths from root to the u_i and $\hat{u}_i$ [Liu *et al.*, 2021b].

For large taxonomies, we adopt the F1 score to evaluate the performance of automatical taxonomies expansion since these taxonomies are DAG-structured rather than tree-structured, i.e., a single node may have multiple parents.

Baselines Comparison

We compare TaxoPrompt with the following baseline taxonomy expansion methods:

- **BERT+MLP** adopts the pre-trained concept embeddings from BERT and leverages a Multi-Layer Perceptron (MLP) for the *is-a* relations identification. The experimental results are from [Yu *et al.*, 2020].
- **TaxoExpan** [Shen *et al.*, 2020] incorporates hierarchical positional information by adopting position-enhanced graph neural networks (GNN). It trains a log-bilinear model to identify a candidate concept.
- **STEAM** [Yu *et al.*, 2020] solves the taxonomy expansion by learning to insert query concepts into mini-paths with a multi-view co-training procedure.
- **TMN** [Zhang *et al.*, 2021] leverages auxiliary and primal signals based on the neural tensor network and regulates concept embeddings via the channel-wise gating mechanism.
- **TEMP** [Liu *et al.*, 2021b] relies on the pre-trained contextual encoder as its core and preserves taxonomical structure information in taxonomy-paths.
- **HEF** [Wang *et al.*, 2021] models the taxonomy with the ego-tree structure to exploit the hierarchical information for taxonomies coherence maintenance.
- **HyperExpan**[Ma *et al.*, 2021] preserves taxonomical structure information in a hyperbolic space. It leverages a hyperbolic graph neural network (HGNN) for encoding concept embedding.

Implementation Details

We use BERT_{base} (uncased) in experiments. The optimizer is AdamW [Loshchilov and Hutter, 2019] with a learning rate of 1e-5. For length κ and times τ of random walk, we set them as 6 and 5 for best performance after the extensive search. We set the batch size to 6 and train the model with 15 epochs. For descriptive context construction, we follow previous work to leverage Wikipedia summary [Liu *et al.*, 2021b] and WordNet definition [Wang *et al.*, 2021]. We use Wikipedia summary for Environment, MAG-CS/PSY, and WordNet definition for WordNet-Noun/Verb. For the rest datasets, we combine Wikipedia summary and WordNet definition.

4.2 Performance Comparison (RQ1)

Table 2 presents overall experimental results on three low-resource taxonomies and Table 3 shows the F1 results on four large taxonomies. We have the following observations:

First, BERT+MLP performs the worst since pre-trained language models are not designed for word-level representations, and such representations provide little contextual information. TaxoExpan propagates the neighborhood information into embeddings via graph neural networks and consistently outperforms the BERT+MLP.

Second, STEAM further improves the performance of TaxoExpan by leveraging mini-paths for hierarchical information capture. TMN formulates the anchor position as a candidate hypernym and hyponym pair, and such a local path structure has been proven effective for leaf expansion.

Third, transformer-based methods like TEMP and HEF achieve state-of-the-art performance and outperform previous methods with a large margin. Their success can be attributed to the better structural information capture and contextual relation extract in the fine-tuning paradigm.

TaxoPrompt consistently outperforms all the baselines on three low-resource benchmarks. Specifically, TaxoPrompt improves state-of-the-arts by 3.7%, 1.5%, and 2.8% for Acc, MRR, and Wu&P on average, confirming the effectiveness of the prompt tuning paradigm for the hypernym generation. Improvement on MRR and Wu&P shows that TaxoPrompt tends to rank the ground truth high and predict semantically similar answers for query concepts. Such improvement relies on that TaxoPrompt can better leverage knowledge in LMs and exploit the structure information of taxonomies compared with all baselines.

Methods	Verb	Noun	PSY	CS
TaxoExpan	12.40	19.90	29.46	19.67
TMN	14.00	20.90	29.11	19.81
HyperExpan	15.00	24.15	32.47	19.92
TaxoPrompt	**25.39**	**41.44**	**33.12**	**21.88**

Table 3: Results of the F1 score on four large datasets (in %). Results of baselines come from [Ma *et al.*, 2021].

#	Setting	Acc	MRR	Wu&P
1	TaxoPrompt	61.4	68.7	85.6
2	w/o two contexts	45.6	54.4	76.2
3	w/o descriptive context	50.0	57.0	76.9
4	w/o taxonomic context	57.5	66.1	83.3
5	#4 + negative sampling	58.7	68.3	84.7
6	#1 - sibling - nephew	59.2	68.0	83.5

Table 4: Ablation studies on science dataset (in %). "w/o" means "without".

Finally, results from Table 3 show that TaxoPrompt automatically expands the four large taxonomies better than the state-of-the-art method HyperExpan. We find TaxoPrompt improves HyperExpan vastly on taxonomies with deeper depth and high-quality descriptive contexts. This observation further demonstrates the ability of TaxoPrompt to better leverage both semantical and structural knowledge.

4.3 Ablation Studies (RQ2)

We conduct experiments on the science dataset for ablation studies and have the following observations:

As shown in Table 4, both descriptive context and taxonomic context contribute to TaxoPrompt (lines 1-4). Compared with line 1, we find Acc, MRR, and Wu&P drop 3.9%, 2.6%, and 2.3% respectively in line 4 without the structure knowledge infused by taxonomic context. To further explore the impact of taxonomic context on distinguishing similar concepts, we replace it with negative sampling in line 5 and train the model with margin ranking loss as in [Liu *et al.*, 2021b]. The results show that our proposed taxonomic context can lead the LM to distinguish negative answers better.

We further study whether the taxonomic context learns global structure information instead of local by restricting random walking areas. In line 6, we forbid nodes from walking to their siblings or uncles, and the constructed taxonomic context is downgraded to separate paths like mini-paths [Yu *et al.*, 2020] or taxonomy-paths [Liu *et al.*, 2021b]. The Acc, MRR, and Wu&P results go down to 59.2%, 68.0%, and 83.5% since the model fails to capture relations between these paths for global structure learning.

4.4 Prompt Discussion (RQ3)

In this section, we discuss the impact of the prompt template and language models on the taxonomy expansion task.

Template	Acc	MRR	Wu&P
[X], [MASK].	58.8	68.3	84.9
[X] is a [MASK].	57.9	67.8	83.2
[MASK], such as [X].	57.0	66.4	83.0
What's parent-of [X]?It's [MASK].	61.4	68.7	85.6

Table 5: Impact of different prompt templates on science dataset (in %).

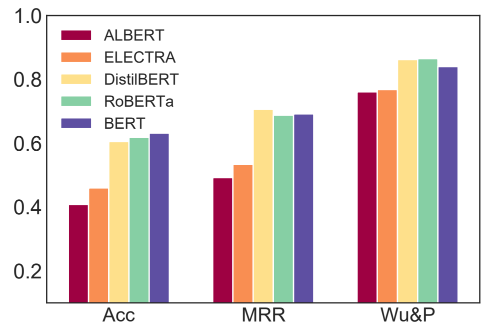

Figure 4: Results of different language models over science dataset. We initialize them using albert-base-v2, roberta-base, electra-small-discriminator and distilbert-base-uncased.

The Effect of Prompt Template. Table 5 shows the experimental results on science dataset using different prompt templates. Designing an appropriate template for the taxonomy expansion task is essential as the Acc difference between the worst and the best template comes to 4.4%. An effective template will help LMs better exploit task-specific knowledge. Besides, prompt tuning can consistently benefit Wu&P under different templates.

The Effect of Language Models. The choice of language models is another key problem for prompt tuning. As shown in Figure 4, DistillBERT [Sanh *et al.*, 2019] achieves the similar performance with BERT. We find that the pre-trained knowledge stored in BERT essentially improves the performance of the hypernym generation. Besides, RoBERTa [Liu *et al.*, 2019] also has remarkable power on the hypernym generation since it exploits dynamic masking for pre-training. We observe that ELECTRA [Clark *et al.*, 2020] fails to achieve the best performance as it did in the fine-tuning solution[Liu *et al.*, 2021b]. One possible reason can be that ELECTRA is pre-trained with the discriminative replaced token detection (RTD) task instead of the MLM task.

5 Conclusions

We propose TaxoPrompt to solve taxonomy expansion efficiently by prompt tuning. TaxoPrompt utilizes a random walk algorithm to capture the global structure of taxonomies and infuses structure knowledge into the LM via taxonomic context. Experimental results show that TaxoPrompt outperforms state-of-the-art methods. Further ablation studies demonstrate the effectiveness of our key designs. In future work, we plan to study the relationship between taxonomic context and negative sampling under the prompt tuning paradigm.

Acknowledgments

This research is supported by Chinese Scientific and Technical Innovation Project 2030 (No.2018AAA0102100), National Natural Science Foundation of China (No.U1936206, 62077031). We thank the reviewers for their constructive comments.

References

[Bordea *et al.*, 2016] Georgeta Bordea, Els Lefever, and Paul Buitelaar. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *SemEval-2016*, pages 1081–1091, 2016.

[Chen *et al.*, 2020] Fenxiao Chen, Yun-Cheng Wang, Bin Wang, and C-C Jay Kuo. Graph representation learning: a survey. *APSIPA*, 9(1):1–21, 2020.

[Clark *et al.*, 2020] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *ICLR*, pages 1–18, 2020.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.

[He *et al.*, 2020] Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. In *EMNLP*, pages 4604–4614, 2020.

[Huang *et al.*, 2019] Jin Huang, Zhaochun Ren, Wayne Xin Zhao, Gaole He, Ji-Rong Wen, and Daxiang Dong. Taxonomy-aware multi-hop reasoning networks for sequential recommendation. In *WSDM*, pages 573–581, 2019.

[Jiang *et al.*, 2020] Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. X-FACTR: multilingual factual knowledge retrieval from pretrained language models. In *EMNLP*, pages 5943–5959, 2020.

[Jurgens and Pilehvar, 2016] David Jurgens and Mohammad Taher Pilehvar. Semeval-2016 task 14: Semantic taxonomy enrichment. In *SemEval-2016*, pages 1092–1102, 2016.

[Karamanolakis *et al.*, 2020] Giannis Karamanolakis, Jun Ma, and Xin Luna Dong. Txtract: Taxonomy-aware knowledge extraction for thousands of product categories. In *ACL*, pages 8489–8502, 2020.

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

[Liu *et al.*, 2021a] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

[Liu *et al.*, 2021b] Zichen Liu, Hongyuan Xu, Yanlong Wen, Ning Jiang, Haiying Wu, and Xiaojie Yuan. TEMP: taxonomy expansion with dynamic margin loss through taxonomy-paths. In *EMNLP*, pages 3854–3863, 2021.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, pages 1–18, 2019.

[Ma *et al.*, 2021] Mingyu Derek Ma, Muhao Chen, Te-Lin Wu, and Nanyun Peng. Hyperexpan: Taxonomy expansion with hyperbolic representation learning. In *EMNLP*, pages 4182–4194, 2021.

[Manzoor *et al.*, 2020] Emaad Manzoor, Rui Li, Dhananjay Shrouty, and Jure Leskovec. Expanding taxonomies with implicit edge semantics. In *WWW*, pages 2044–2054, 2020.

[Nickel and Kiela, 2017] Maximilian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *NeurIPS*, pages 6338–6347, 2017.

[Sanh *et al.*, 2019] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[Schick and Schütze, 2021] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In *EACL*, pages 255–269, 2021.

[Schuster and Nakajima, 2012] Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *ICASSP*, pages 5149–5152, 2012.

[Shen *et al.*, 2018] Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T Vanni, Brian M Sadler, and Jiawei Han. Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion. In *SIGKDD*, pages 2180–2189, 2018.

[Shen *et al.*, 2020] Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. Taxoexpan: Self-supervised taxonomy expansion with position-enhanced graph neural network. In *WWW*, pages 486–497, 2020.

[Sinha *et al.*, 2015] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. An overview of microsoft academic service (MAS) and applications. In *WWW*, pages 243–246, 2015.

[Song *et al.*, 2021] Xiangchen Song, Jiaming Shen, Jieyu Zhang, and Jiawei Han. Who should go first? a self-supervised concept sorting model for improving taxonomy expansion. *arXiv preprint arXiv:2104.03682*, 2021.

[Takeoka *et al.*, 2021] Kunihiro Takeoka, Kosuke Akimoto, and Masafumi Oyamada. Low-resource taxonomy enrichment with pretrained language models. In *EMNLP*, pages 2747–2758, 2021.

[Wang *et al.*, 2021] Suyuchen Wang, Ruihui Zhao, Xi Chen, Yefeng Zheng, and Bang Liu. Enquire one's parent and child before decision: Fully exploit hierarchical structure for self-supervised taxonomy expansion. In *WWW*, pages 3291–3304, 2021.

[Yang *et al.*, 2017] Shuo Yang, Lei Zou, Zhongyuan Wang, Jun Yan, and Ji-Rong Wen. Efficiently answering technical questions—a knowledge graph approach. In *AAAI*, pages 3111–3118, 2017.

[Yu *et al.*, 2020] Yue Yu, Yinghao Li, Jiaming Shen, Hao Feng, Jimeng Sun, and Chao Zhang. STEAM: self-supervised taxonomy expansion with mini-paths. In *KDD*, pages 1026–1035, 2020.

[Zeng *et al.*, 2021] Qingkai Zeng, Jinfeng Lin, Wenhao Yu, Jane Cleland-Huang, and Meng Jiang. Enhancing taxonomy completion with concept generation via fusing relational representations. In *KDD*, pages 2104–2113, 2021.

[Zhang *et al.*, 2021] Jieyu Zhang, Xiangchen Song, Ying Zeng, Jiaze Chen, Jiaming Shen, Yuning Mao, and Lei Li. Taxonomy completion via triplet matching network. In *AAAI*, pages 4662–4670, 2021.