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

Hongyuan Xu¹,², Yunong Chen¹,², Zichen Liu¹,², Yanlong Wen¹,²* and Xiaojie Yuan¹,²

¹College of Computer Science, Nankai University
²Tianjin 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.

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

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.
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 ego-net 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 \( u \in \mathcal{N} \) represents a concept (i.e., a word or a phrase) and each directed edge \((u, v) \in \mathcal{E}\) 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 \( \mathcal{T}^0 = (\mathcal{N}^0, \mathcal{E}^0) \) 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 \( \mathcal{T} = (\mathcal{N}^0 \cup \mathcal{C}, \mathcal{E}^0 \cup \mathcal{R}) \) 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 \( \mathcal{N} \) and (2) a new concept \( c \in \mathcal{C} \), our goal is to generate a list of tokens \( \mathcal{L} = (\ell_1, \ell_2, \ldots, \ell_|\mathcal{L}|) \), 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 \((u, c)\) to the existing taxonomy. Mathematically, our final taxonomy expansion goal can be formulated as following \(|\mathcal{C}|\) independent optimization problems [Shen et al., 2020]:

\[
\hat{u}_i = \arg \max_{u_i \in \mathcal{N}^0} \log P(c_i | u_i, \Theta), \forall i \in \{1, 2, \ldots, |\mathcal{C}|\},
\]

where \( \Theta \) 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\textsubscript{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 \( c \):

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

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:

\[
s = t_1, t_2, \ldots, t_i, \langle \text{mask} \rangle_1, \ldots, \langle \text{mask} \rangle_{|\mathcal{L}|}, t_j, \ldots, t_n,
\]
Where $t_{i,j}$ is the left-to-right context for masked positions and $t_{j,m}$ is the opposite. Finally, the LM is applied for parallel masked positions prediction by calculating the conditional probability distribution [Jiang et al., 2020]:

$$
\hat{t}_k = \arg\max_{t'_k \in V} P(t'_k | t_1, \ldots, t_{k-1}, t_{k+1}, \ldots, t_n, \Theta),
$$

(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:

$$
\begin{align*}
O &= \text{LayerNorm}(\sigma(HD^T)) \\
L &= OM^T
\end{align*}
$$

(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 $M \in \mathbb{R}^{|V| \times h}$ are learnable projection matrices. $\sigma$ represents the activation function. $L \in \mathbb{R}^{n \times |V|}$ 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:

$$
\hat{t}_k = \arg\max_{t'_k \in V} L(k, t'_k),
$$

(5)

after $|L|$ times prediction, the token list $L$ 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.

**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 $R$, where $R = \{ \text{parent-of, child-of, sibling-of, nephew-of, posterity-of} \}$. Then, we design a three-stage random walk algorithm: (1) Selecting a random relation $r_i$ from $R$, 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 $k$ times random walk, we serialize the taxonomy into a walked path named taxonomic context (shown in Figure 3):

$$
W(c) = c, r_1, u_1, r_2, \ldots, u_{i-1}, r_i, u_i, \ldots, u_{k-1}, r_K, u_K,
$$

(6)

where $W(c)$ denotes one taxonomic context for a query concept $c \in N^0$ and $k$ is a hyperparameter. Finally, we concatenate $\tau$ taxonomic contexts to construct the full taxonomic context $W'(c)$. 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 $D(c)$. [He et al., 2020] has demonstrated that the LM can be complemented by $D(c)$ since the LM learns to summarize the main attributes from the description:

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

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:

\[
P(c) \Pr[SEP|D(c)], \quad P(c) \Pr[SEP|W'(c)]
\]

and the former prompt is applied during inference.

3.4 Learning and Inference

Training Data Construction. Given one edge \((u, c)\) from the existing taxonomy \(T^0 = (\mathcal{N}^0, \mathcal{E}^0)\), 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 \(L_{\text{gold}}\) is constructed by tokenizing the parent node \(u\). Notice that the number of masked tokens in \(s\) equals to \(|L_{\text{gold}}|\). Finally, one training instance \(X = \langle s, L_{\text{gold}} \rangle\) corresponds to the edge \((u, c)\) is created. By repeating the above process for each edge in \(T^0\), we obtain the full training dataset \(X = \{X_1, X_2, \ldots, X_{|\mathcal{E}^0|}\}\).

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

\[
\mathcal{L}(\Theta) = - \frac{1}{|X|} \sum_{X \in X} \log \left( \sum_{t^* \in L_{\text{gold}}} \exp \left( L(k, t^*) \right) \right),
\]

where \(t^*\) 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 a 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:

\[
score(u, c) = \frac{1}{|L_u|} \sum_{\ell_i \in L_u} L(k, \ell_i),
\]

where \(L_u = \text{tokenize}(u)\) and \(\ell_i\) represents \(i\)-th token in \(L_u\). \(k\)-th position of \(s\) is the \(i\)-th mask token, where \(\ell_i\) is supposed to be filled in.

Complexity Analysis. The time complexity of training is \(O(I \cdot |\mathcal{E}^0| \cdot t_{\text{avg}} \cdot d)\), where \(I\) is the number of iterations, \(t_{\text{avg}}\) 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 \(O(|\mathcal{C}| \cdot t_{\text{avg}}^2 \cdot d)\), where \(|\mathcal{C}|\) 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 \(O(|\mathcal{C}|^2 \cdot t_{\text{avg}}^2 \cdot d)\).

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)?

4.1 Experimental Setup

Datasets

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

| Dataset          | \(|\mathcal{N}|\) | \(|\mathcal{E}|\) | \(|D|\) |
|------------------|------------------|------------------|--------|
| 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. \(|D|\) indicates the taxonomy depth.

**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:

\[
\text{MRR} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\text{rank}(u_i)}
\]
### 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 $k$ and times $\tau$ 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 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.

![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.](image-url)
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.

---

<table>
<thead>
<tr>
<th>Methods</th>
<th>Verb</th>
<th>Noun</th>
<th>PSY</th>
<th>CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>TaxoExpan</td>
<td>12.40</td>
<td>19.90</td>
<td>29.46</td>
<td>19.67</td>
</tr>
<tr>
<td>TMN</td>
<td>14.00</td>
<td>20.90</td>
<td>29.11</td>
<td>19.81</td>
</tr>
<tr>
<td>HyperExpan</td>
<td>15.00</td>
<td>24.15</td>
<td>32.47</td>
<td>19.92</td>
</tr>
<tr>
<td>TaxoPrompt</td>
<td>25.39</td>
<td>41.44</td>
<td>33.12</td>
<td>21.88</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>#</th>
<th>Setting</th>
<th>Acc</th>
<th>MRR</th>
<th>Wu&amp;P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TaxoPrompt</td>
<td>64.1</td>
<td>68.7</td>
<td>85.6</td>
</tr>
<tr>
<td>2</td>
<td>w/o two contexts</td>
<td>45.6</td>
<td>54.4</td>
<td>76.2</td>
</tr>
<tr>
<td>3</td>
<td>w/o descriptive context</td>
<td>50.0</td>
<td>57.0</td>
<td>76.9</td>
</tr>
<tr>
<td>4</td>
<td>w/o taxonomic context</td>
<td>57.5</td>
<td>66.1</td>
<td>83.3</td>
</tr>
<tr>
<td>5</td>
<td>#4 + negative sampling</td>
<td>58.7</td>
<td>68.3</td>
<td>84.7</td>
</tr>
<tr>
<td>6</td>
<td>#1 - sibling - nephew</td>
<td>59.2</td>
<td>68.0</td>
<td>83.5</td>
</tr>
</tbody>
</table>

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

---

**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, DistilBERT [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.

---

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

<table>
<thead>
<tr>
<th>Template</th>
<th>Acc</th>
<th>MRR</th>
<th>Wu&amp;P</th>
</tr>
</thead>
<tbody>
<tr>
<td>[X], [MASK].</td>
<td>58.8</td>
<td>68.3</td>
<td>84.9</td>
</tr>
<tr>
<td>[X] is a [MASK].</td>
<td>57.9</td>
<td>67.8</td>
<td>83.2</td>
</tr>
<tr>
<td>[MASK], such as [X].</td>
<td>57.0</td>
<td>66.4</td>
<td>83.0</td>
</tr>
<tr>
<td>What’s parent-of [X]?It’s [MASK].</td>
<td>61.4</td>
<td>68.7</td>
<td>85.6</td>
</tr>
</tbody>
</table>

---

**Table 4:** Results of the F1 score on four large datasets (in %). Results of baselines come from [Ma et al., 2021].
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


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


