Записки научных
семинаров ПОМИ
Том 546, 2025 г.

F. Sadkovskii, N. Loukachevitch, I. Grishin

FINE-TUNING LARGE LANGUAGE MODELS FOR HYPERNYM DISCOVERY TASK: SISTER TERMS DO THEIR PART

Abstract. This paper addresses the problem of bias introduced by cohyponyms–nodes sharing the same parent hypernym–in training datasets for hypernym discovery tasks. While the removal of test items from training data is essential for preventing data leakage, we argue that excluding cohyponyms is equally critical. When fine-tuning a model on on a dataset composed of hyponym-hypernym pairs extracted from a taxonomic resource WordNet, puncturing only test nodes is not enough to adequately assess the quality of the model on test data. Cohyponyms act as implicit hints for identifying hypernyms, artificially enhancing the performance of model and misrepresenting its utility in real-world scenarios. We fine-tuned LLaMA-2 using the TaxoLLaMA training procedure of Moskvoretskii et al. (2024) on an extensive number of WordNet-derived subsamples of hyponym–hypernym pairs with and without their definitions. Evaluation on the SemEval-2018 dataset showed that including co-hyponyms in the training data artificially inflates performance metrics.

§1. Introduction

Large language models (LLMs) have revolutionized natural language processing, driving significant advances in various natural language processing tasks. These models also hold immense promise for taxonomy enrichment – the expansion and refinement of hierarchical knowledge structures crucial for information science applications like search engines, recommendation systems, and content categorization. Maintaining and enriching taxonomies manually is an incredibly resource-intensive task, especially in the face of the exponential growth of available information. Given the importance of taxonomy replenishment, various approaches have been

Key words and phrases: Hypernym Discovery, Taxonomy Enrichment, WordNet, TaxoLLaMA..

This work was supported by the Interdisciplinary Scientific and Pedagogical School of Lomonosov Moscow State University (grant No. 23-ShCh05-11) within the state assignment (registration No. 124020100068-4).

developed over time, such as leveraging text patterns [1,16,26,32–35], vector representations [5,12,28,39,42] or both [3,7,45]. More recently, with the rise of LLMs new methods based on masked word predictions [15, 44, 46] and generative models [2,21,29,40] have emerged. LLMs, with their ability to detect complex semantic relationships, generate contextually appropriate terms, and identify meaningful connections within vast text corpora, provide a powerful solution for automating critical aspects of this process. This includes adding and refining categories as well as ensuring consistency across diverse knowledge domains.

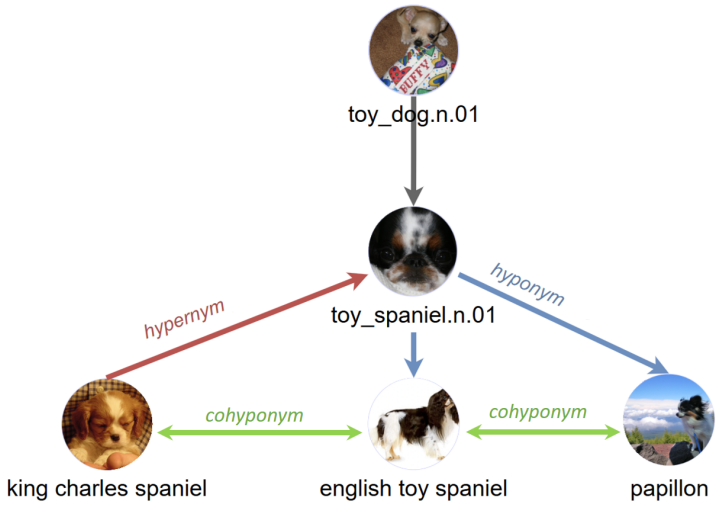

Figure 1. An example of a taxonomic subgraph with relations between the nodes, adapted from [29]

A promising recent approach, (TaxoLLaMA) [24,25], leverages this potential. The authors fine-tuned a non-instructive LLaMA-2 7B model [41] on hyponym-hypernym pairs, extracted from WordNet [23] taxonomy. This resulted in state-of-the-art performance in taxonomy enrichment and related tasks like hypernym extraction, taxonomy construction, and lexical entailment after further fine-tuning on specific evaluation datasets.

However, utilizing a pre-existing resource like WordNet for hypernym discovery requires careful preparation of the training data. This is because

the dataset may contain inadvertent cues—such as test items or related terms—that could artificially inflate performance. Specifically, words for which a hypernym is to be predicted should be excluded from the training set. The authors of TaxoLLaMA accounted for this concern and released a special version of their model for evaluation purposes. In this version, a "cleansed" training dataset was used, ensuring that none of the test set items were present in training.

While we acknowledge the importance of removing test items, we argue that this step alone is insufficient. Cohyponyms (i.e. nodes with the same parent hypernym, see Fig. 1) of the target words should also be excluded from the training data, as their presence in the training sample introduces significant bias on evaluation. Cohyponyms act as implicit hints for identifying the correct hypernym, which can lead to an overestimation of the performance of the model. Addressing this issue is essential because, in real-world applications of taxonomy enrichment, developers often encounter scenarios where cohyponyms of newly introduced concepts are unavailable in the training data. Thus, unbiased results, free from this cohyponym factor, are more representative of a model's practical utility.

In this paper we will show that the presence of cohyponyms in the training dataset increases model performance and therefore constitutes a form of data leakage. We focused on Hypernym Discovery task and conducted several experiments on SemEval-2018 "1A: English" dataset [10]. The task was chosen due to its fundamental importance in solving other tasks related to taxonomic relationships.

§2. Related Work

Research at the intersection of Taxonomies and Language Models has historically been dominated by encoder-based architectures rather than GPT-style models. Notable contributions in this field include the CTP (**c**onstructs **t**axonomic trees using **p**retrained language models) approach [11], in which the authors took advantage of language models like BERT [14] and RoBERTa [22]. In [15], a similar approach based on template prompts and a BERT model for mask filling is proposed. The model was tested on both BLESS-like [4] and SemEval-2018 datasets [10].

At the same time, there has been limited research comparing these approaches with more recent decoder-type open-source models such as LLaMA-2 [41] and Mistral [19] for taxonomy-related tasks. Filling this gap is also one the main goals of the authors of TaxoLLaMA.

The Hypernym Discovery task, which involves generating hypernyms for given hyponyms, has seen significant developments. A major advancement came from [29], who introduced a taxonomy-adapted, fine-tuned T5 model [13]. Their research demonstrated that encoder-decoder models, particularly the T5 series, outperformed encoder and decoder-only architectures. Their approach incorporated both zero-shot and few-shot methodologies, utilizing template-based prompts and developing specialized prompt formats for hypernym prediction and taxonomy representation.

Several earlier approaches made important contributions to the field. These include the 300-sparsans method [6], which enhanced traditional word2vec techniques; the Hybrid model [17], combining k-Nearest Neighbor with Hearst patterns; and CRIM [7], the top performer in the SemEval-2018 competition, utilizing a Multilayer Perceptron (MLP) with contrastive loss and Hearst patterns for both hyponym and hypernym extraction. The Recurrent Mapping Model (RMM) [3] further advanced the field with its MLP architecture incorporating residual connections.

Specialized models for hypernym prediction have also emerged, such as Hypert [46], a BERT-based encoder model that predicts hyponym-hypernym relationships using projection matrix learning. While Hypert achieved superior results compared to SemEval-2018 Task 9 competitors like CRIM, its computational intensity (766 times longer processing time than standard BERT) presents practical limitations.

In the era of distributional semantics, it has been observed that hypernym discovery using vector similarity is challenging because cohyponyms, in addition to hypernyms, are among the most similar terms in the vector space. Weeds et al. [43] demonstrated that distinguishing between these types of semantic relations is difficult using unsupervised distributional methods. The authors achieved a significant improvement by employing supervised SVM training on both positive and negative examples.

Several studies have demonstrated that augmenting the list of predictions for input terms with predictions for their cohyponyms enhances results by achieving higher recall. The authors of [38] and [31] demonstrated the limitations of using Hearst patterns exclusively for hypernym identification. More recent approaches incorporate Hearst patterns specifically designed to identify cohyponyms: the authors of [7] employed enumeration patterns to search for cohyponyms in textual corpora, while in [40] multiple patterns to generate cohyponyms with decoder-type LLMs were

utilized. These findings suggest that cohyponyms possess distinctive information content that differentiates them from other items within taxonomic hierarchies.

§3. Model Description

To assess the hypothesis that the presence of cohyponyms of the test items in the training set enhances the performance of model, we fine-tune LLaMA-2 7B[1] [41] on specifically constructed training samples using the TaxoLLaMA methodology as a foundation.

TaxoLLaMA is a LLM-based approach to various taxonomy-related tasks, first introduced by [24] and further developed in [25]. The authors describe it as "the everything-in-one model" due to its versatility in handling multiple tasks simultaneously, including Taxonomy Enrichment, Hypernym Discovery, Taxonomy Construction, and Lexical Entailment. The model is available in two variants: *TaxoLLaMA*, which was fully trained, and TaxoLLaMA$_{\text{bench}}$, a version specifically designed to ensure that no test items were included in the training data, allowing for accurate evaluation on benchmark tests. TaxoLLaMA$_{\text{bench}}$ achieved impressive results, setting 11 state-of-the-art (SoTA) results and securing 4 second-place positions out of 16 tasks on the benchmark.

The model offers several notable advantages, including a relatively small weight thanks to LoRA implementation and 4-bit quantization, as well as strong zero-shot performance on taxonomy-related tasks. Its performance can be significantly enhanced through fine-tuning on benchmark datasets, such as the SemEval-2018 datasets [10]. Fig. 2 illustrates the pipeline utilizing the "1A: English" dataset as an exemplar. After such fine-tuning, TaxoLLaMA achieved a Mean Reciprocal Rank (MRR) value exceeding 50, substantially outperforming the previous SoTA of 45 reported by [29].

The authors conducted an ablation study revealing that the most effective fine-tuning approach for Taxonomy Enrichment and Hypernym Discovery tasks involved predicting a single hypernym from a hyponym along with its WordNet definition. They employed a specific prompt structure (where (1) is the system prompt, (2) is the varying instruction and (3) is an eventual answer of the model) [24, p. 3]:

> (1) [INST] <SYS> You are a helpful assistant. List all the possible words divided with a comma. Your answer should

[1] https://huggingface.co/meta-llama/Llama-2-7b-hf

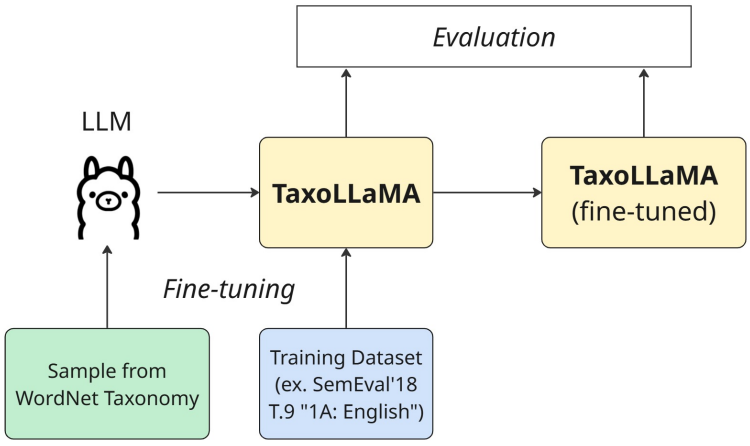

Figure 2. The TaxoLLaMA pipeline comprising two fine-tuning stages: primary fine-tuning on hyponym–hypernym pairs from WordNet, followed by secondary fine-tuning on a task-specific dataset such as "1A: English". Evaluation metrics are applied after both stages.

not include anything except the words divided by a comma </SYS>

(2) hyponym: tiger (large feline of forests in most of Asia having a tawny coat with black stripes) | hypernym: [/INST]

(3) big cat,

The authors have made their datasets, model, and code publicly available[2]. We utilized this code in our research for fine-tuning LLaMA-2 model, with minor modifications implemented to create experimental dataset variants, see section 4.

§4. Dataset

Following the main idea of TaxoLLaMA approach, we used two types of datasets in our experiments.

[2] https://github.com/VityaVitalich/TaxoLLaMA

First, for evaluation and additional fine-tuning we used SemEval-2018 Task 9 "1A: English" dataset[3]. This dataset was chosen primarily to maintain consistency with the established practice of using it as a benchmark [3, 15, 24, 25, 29], enabling comparisons with prior approaches. It is also crucial since we use TaxoLLaMA$_{\text{bench}}$'s performance on this dataset as a baseline. By focusing on a single dataset, we can more effectively track the relationship between the sample content and the model's output for that specific dataset.

The dataset contains elements representing concepts and entities from various subject areas, each accompanied by a list of gold hypernyms. The authors of the competition task collected the hyponyms from the large text corpos and provided the list of ground truth hypernyms, extracted from multiple sources, with a primary focus on WordNet and the English-language Wiki system. A few examples taken from the dataset are presented in Table 1.

split	*data*	*gold*
training	blackfly	homopterous insect, insect
training	Turonian	technical specification, geologic timescale, physical property, geological period, magnitude, unit of time, geological time, geologic time
training	tropical storm	atmosphere, windstorm, violent storm, air current, atmospheric state, density, current of air, storm damage, atmospheric phenomenon, storm, cyclone, natural phenomenon, tempest, wind
test	maliciousness	malevolence, distaste, hatred, hate, malignity
test	quo warranto	proceedings, legal proceedings, proceeding, due process, legal proceeding
test	Jeff Francis	thrower, baseball, player, jock, person

Table 1. Sample items from "1A: English" dataset.

The dataset consists of the two splits: 'training' and 'test', both containing 1500 elements. We use the training split for fine-tuning and the test split for evaluation.

Second, to conduct the initial fine-tuning, we collected several subsamples from WordNet 3.0 using the Python NLTK package [8], following the

[3]The data is available at `https://drive.google.com/file/d/14_RgB3_it7a_1mLXeRCyzwY5BHdWgnlP/view`

algorithm proposed in [25][4]. In [25], the authors expand on their prior work [24] and offer a more comprehensive explanation of their methodology. They thoroughly describe the data collection process and the algorithm used to compile the dataset for fine-tuning the LLaMA-2 model. Specifically, the authors built a directed acyclic graph from the WordNet synsets, iteratively adding items and linking them to their hypernyms. Then they extracted available hyponym–hypernym pairs from the graph.

The NLTK WordNet 3.0 implementation comprises 82,115 noun synsets that are organized into 20 topological generations. Excluding the most general term, "entity", the first generation (i.e., the set of nodes without parent nodes) contains an additional 7,725 synsets. These synsets predominantly represent specific terms from domains such as geography, politics, and religion. The unusually large number of synsets lacking assigned hypernyms can be attributed to the manual nature of the WordNet annotation process, which introduces inconsistencies and deviations from the principle of coherence.

To ensure the efficiency of the dataset, only pairs in which the hypernym is not a top-level node in the (sub)graph and the hyponym is unambiguous were included. For the $\text{TaxoLLaMA}_{\text{bench}}$ training, the authors meticulously excluded any items present in the benchmark test datasets, namely SemEval-2018 Task 9 [10], TexEval-2 [9] and MAGs [37] from the final version of their dataset. Initially, the dataset contained 44,772 pairs, but after removing items that overlapped with the test datasets, this was reduced to 36,755 pairs. Definitions for the hyponyms were straightforwardly extracted from WordNet.

In our study, we began by assembling two base training datasets—one with definitions and one without—following the approach of Moskvoretskii et al. to training $\text{TaxoLLaMA}_{\text{bench}}$. From these base samples, we generated 66 modified training samples by either reducing the number of specific items in the sample without definitions or omitting the definitions in the settings with definitions.

The primary distinction of our approach lies in the handling of test nodes; unlike the previous approach, we did not include test items from all three referenced datasets. Instead, since our focus was on the Hypernym Discovery task and we selected "1A: English" as the benchmark, we excluded only the test items from this particular dataset. This adjustment

[4]The samples we used for training and the code to collect them is available at https://github.com/feudor2/TaxoLLaMA

allowed us to slightly expand the base training sample to 40,636 pairs after removal of the test items.

The modification process entailed adjusting the cleansed dataset by varying the types of nodes targeted for modification and controlling the percentages of these nodes.

Specifically, the modifications were implemented through three approaches:

(1) removing nodes of a certain type;
(2) partially adding definitions to nodes of a certain type;
(3) partially excluding definitions from the sample where definitions were initially provided.

The target nodes were categorized into three types: cohyponyms only, non-cohyponyms only, and random nodes. The modifications were applied at different levels: 0%, 25%, 50%, 75%, and 100%, with 100% representing the total number of cohyponyms of the removed test nodes that occur as hyponyms in our base training sample. This number is 5,535 unique synsets (12.37% of all synsets), while the total number of dataset items that include these nodes is 8,776 (19.66% of all items).

In the removal setting, we selected a fixed random seed for sampling cohyponyms and non-cohyponyms, but employed three different seeds for obtaining random samples. For both cohyponym and non-cohyponym conditions, we conducted a 4-fold cross-validation at 25%, 50%, and 75% removal rates. The cross-validation process involved the following steps: we partitioned the total number of nodes of both types into four equal parts. In the cohyponym condition, partitions were normalized based on the number of hyponym-hypernym pairs that involve cohyponym nodes. For the non-cohyponym condition, normalization was based on the shortest path from non-cohyponym nodes to the nearest test item within the complete WordNet graph. After normalization, we adjusted the partition sizes to match those in the cohyponym condition.

To control the proportions of removed items, we adopted the following approach: for 25% and 75% removal rates, we sequentially removed one or three partitions, respectively, yielding four samples for each proportion. For the 50% removal rate, we removed all six possible combinations of two partitions, resulting in six samples. Additionally, in the random sampling condition, we ensured the presence of cohyponyms in the partitions for removal by setting the probability of selecting a node proportional to the ratio of cohyponym nodes to the total number of nodes.

Table 2 provides detailed information about the sample sizes and the actual percentage of cohyponyms retained after removing nodes.

proportion	actual % of removed cohyponyms by relation type			Size of
	cohyponyms	non-cohyponyms	random	**samples**
25%	24.85±0.47	0	4.11±0.02	38442
50%	50.1±0.37	0	7.84±0.43	36248
75%	74.85±0.43	0	11.57±0.29	34054
100%	100	0	15.36±0.19	31860

Table 2. Sample sizes and proportions of deleted nodes averaged across all folds and random seeds (in percent).

For non-cohyponyms and random settings, the sample sizes were aligned with their corresponding cohyponym samples to ensure a fair comparison. It should be noted that for cohyponyms, the actual shares retained in the dataset were slightly lower than the target percentages. This discrepancy occurred because removing certain nodes caused breaks in the graph structure, resulting in additional losses. To account for this, we also include in the table the actual percentage of removed cohyponyms within each sample.

The modification of definitions was carried out in a similar manner, but with one additional parameter. This parameter determined whether definitions were added to a specific type of node in a sample that typically lacked definitions (the first setting), or whether a certain percentage of definitions was removed from a sample that initially contained them (the second setting). Tables 3 and 4 provide an overview of the samples with the corresponding node types and percentages for the first and the second settings, respectively. For definition settings we fixed only one seed for random samples, which gave us 16 training samples in each condition.

All factors considered in the sample construction procedure are depicted in Fig. 3.

§5. Experiments

5.1. Metrics. For the evaluation of our model, we utilized three standard metrics commonly used in hypernym discovery: Mean Reciprocal Rank (MRR, 1) and Mean Average Precision (MAP, 2). We calculate all the metrics for the 15 first predictions, following SemEval-2018 Task 9

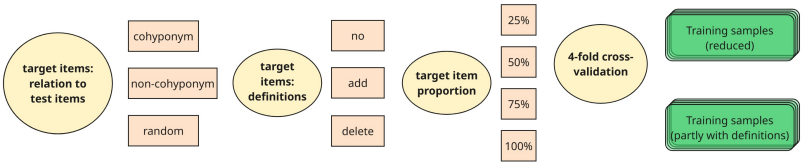

Figure 3. Schematic representation of the key parameters incorporated in the construction of training samples.

percentage of pairs with excluded definitions	*percentage of pairs with definition*	*actual percentage of cohyponyms with definition*
base sample, 0%	0	0
cohyponyms, 25%	3.53	25
cohyponyms, 50%	7.08	50
cohyponyms, 75%	10.6	75
cohyponyms, 100%	14.14	100
non-cohyponyms, 25%	3.55	0
non-cohyponyms, 50%	7.08	0
non-cohyponyms, 75%	10.6	0
non-cohyponyms, 100%	14.05	0
random nodes, 25%	3.55	3.58
random nodes, 50%	7.08	6.97
random nodes, 75%	10.64	10.17
random nodes, 100%	14.15	13.68

Table 3. Sample with definitions with percentage of *included* definitions (1st setting).

evaluation methodology [10].

$$MRR@K = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{rank_i} \tag{1}$$

where N is number of items in the dataset, $rank_i$ is the rank of the first relevant term from the list of predicted hypernyms and K is the maximum size of the list of hypernyms.

$$MAP@K = \frac{1}{N}\sum_{n=1}^{N} AP@K_n \tag{2}$$

percentage of pairs with excluded definitions	*percentage of pairs with definition*	*actual percentage of cohyponyms with definition*
base sample, 0%	100	100
cohyponyms, 25%	96.47	75
cohyponyms, 50%	92.92	50
cohyponyms, 75%	89.39	25
cohyponyms, 100%	85.86	0
non-cohyponyms, 25%	96.45	100
non-cohyponyms, 50%	92.92	100
non-cohyponyms, 75%	89.4	100
non-cohyponyms, 100%	85.85	100
random nodes, 25%	96.45	96.42
random nodes, 50%	92.93	93.03
random nodes, 75%	89.36	89.83
random nodes, 100%	85.85	86.32

Table 4. Sample with definitions with percentage of *excluded* definitions (2nd setting).

where N is the number of items in the dataset and $AP@K_n$ is the average precision on K elements for list n (3).

$$AP@K = \frac{1}{R}\sum_{k=1}^{K}(Precision@k \cdot I[y_k = 1]) \tag{3}$$

where R is the number of ground truth hypernyms for an item, $Precision@k$ is precision at rank k (5.1), K is the maximum number of predicted hypernyms, I is the indicator function and y_i is the label of element i

$$Precision@k = \frac{p}{k} \tag{4}$$

where p is the number of relevant items in the first k elements.

5.2. Experiment Setup. First and foremost, we sought to replicate the original $\text{TaxoLLaMA}_{\text{bench}}$ training pipeline as closely as possible, constructing the same training dataset as described by the authors. This allowed us to assess the reproducibility of their approach. Following this, we trained the model using our own training samples, which were assembled according to the method described in the previous section.

Our experiments involved two fine-tuning procedures, with the model being evaluated on the "1A: English" test dataset after each procedure. All experiments were performed using a single NVIDIA Tesla A100 GPU, with a learning rate of $3 \cdot 10^{-4}$ over one training epoch in a zero-shot setting (i.e., no demonstration examples were provided in the prompt), low-rank adaptation (α=16, r=8, dropout=0.05) and 4-bit quantization.

For the first fine-tuning procedure, we used a batch size of 64 and trained the LLaMA-2 model on one of the dataset samples described in Tab. 2, 3 and 4. The prompt was provided as we showed in section 3, either with or without a definition depending on the experimental setting and type of training sample. In the second fine-tuning procedure, we reduced the batch size to 2 and fine-tuned the model using the "1A: English" training dataset. No definitions were used since they were not provided with the SemEval-2018 Task 9 datasets.

For model inference, we used a batch size of 16 and followed the generation parameters recommended in [25], including the use of 3 beams for beam search, a temperature of 0.8, a maximum of 32 new tokens, a top-k value of 40, and constraints to prevent repeating n-grams with length 3.

5.3. Baselines. We used two implementations of $\text{TaxoLLaMA}_{\text{bench}}$ as baselines for our evaluation: one fine-tuned exclusively on WordNet achieved an MRR score of 37.66, and the other subsequently fine-tuned on the "1A: English" training dataset attained an improved MRR score of 51.59 on the "1A: English" test set [25, p. 10]. But since we narrowed our research to evaluation only on one dataset, and thus use a larger base dataset with less test items removed, we were prompted to train our own baseline $\text{TaxoLLaMA}_{\text{bench}}$-like models on two versions of our dataset — one including definitions and one without them. We refer to these models as $\text{TaxoLLaMA}_{\text{SE(1A)}}$ and $\text{TaxoLLaMA}_{\text{SE(1A)-def}}$, respectively.

§6. Results

This section is organized as follows. In the first subsection, we compare the performance of the models presented in [25] with the reproduced versions, alongside the MRR scores of our own $\text{TaxoLLaMA}_{\text{SE(1A)}}$ variations. The second subsection examines the performance of models trained on 42 samples extracted from WordNet, where no definitions were included in the prompt. In the third subsection, we analyze the results of training models on 24 WordNet samples that included definitions. Finally, in the

fourth subsection, we evaluate the performance of the 66 models discussed in the second subsection after further fine-tuning using the "1A: English" training dataset.

6.1. Baselines. The MRR scores obtained in [25] with those reproduced by us on "1A: English" test dataset are shown in Table 5. "(fine-tuned)" indicates models that were fine-tuned on the training portion of the "1A: English" dataset.

baseline	*MRR*
$\text{TaxoLLaMA}_{\text{bench}}$ [25]	37.66
$\text{TaxoLLaMA}_{\text{bench}}$ (fine-tuned) [25]	51.59
$\text{TaxoLLaMA}_{\text{bench}}$	38.82
$\text{TaxoLLaMA}_{\text{bench}}$ (fine-tuned)	50.11
$\text{TaxoLLaMA}_{\text{bench-def}}$	40.36
$\text{TaxoLLaMA}_{\text{bench-def}}$ (fine-tuned)	**51.63**

Table 5. Baseline models' performance assessed with MRR.

We successfully replicated the metric scores reported for TaxoLLaMA. During the initial fine-tuning, we achieved slightly higher scores compared to the original results, both with and without the use of definitions. Additionally, we validated the authors' hypothesis that training with definitions leads to improved model performance.

In Tab. 6 we provide $\text{TaxoLLaMA}_{\text{SE(1A)}}$ model MRR scores fine-tuned on the base dataset with excluded "1A: English" test items.

baseline	*MRR*
$\text{TaxoLLaMA}_{\text{SE(1A)}}$	39.31
$\text{TaxoLLaMA}_{\text{SE(1A)}}$ (fine-tuned)	**52.21**
$\text{TaxoLLaMA}_{\text{SE(1A)-def}}$	36.66
$\text{TaxoLLaMA}_{\text{SE(1A)-def}}$ (fine-tuned)	49.25

Table 6. $\text{TaxoLLaMA}_{\text{SE(1A)}}$ performance of models assessed with MRR.

By removing only the "1A: English" test items from the WordNet graph, we achieve even better results following both the initial and "1A: English"

dataset fine-tuning. However, a notable contrast emerged when training with and without definitions: while increasing the sample size improves model performance compared to previous baseline experiments, adding more definitions, conversely, reduces it.

The TaxoLLaMA$_{\text{SE(1A)}}$ model variations' performance are referred to as "base samples" for the corresponding experimental setups. For example, "TaxoLLaMA$_{\text{SE(1A)}}$" serves as the baseline model for the initial fine-tuning using samples from WordNet, as well as for fine-tuning using samples that include definitions (the first setting). Meanwhile, "TaxoLLaMA$_{\text{SE(1A)-def}}$ (fine-tuned)" acts as the baseline model for further "1A: English" training set fine-tuning of models that were previously trained on samples where we controlled the elimination of definitions (the second setting).

6.2. Fine-tuning on WordNet. We compare the results of models after fine-tuning on samples with removed nodes introduced in Tab. 2 in Tab. 7 below; the changes of MRR[5] scores across different proportions are depicted in Fig. 4.

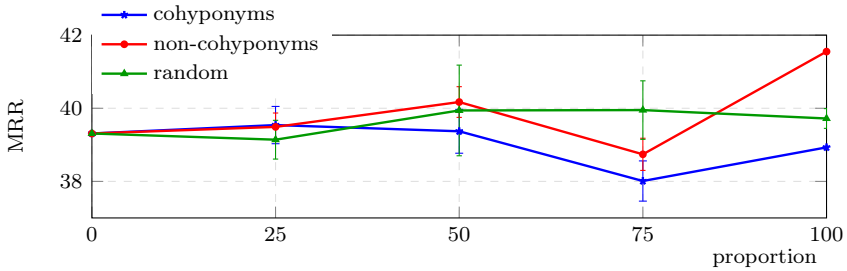

Figure 4. MRR across three conditions in the removal setting.

The performance of model exhibited lower scores when cohyponyms were removed compared to when they were retained across all conditions, with the exception of the 25% node removal condition. In the non-cohyponym condition, the maximum metric score was observed at 100% proportion, while the cohyponym condition reached its peak performance

[5]Henceforth, we report only MRR scores, as they demonstrated a strong positive correlation with MAP scores in our experiments ($r = 0.82$).

percentage of deleted nodes (by number of cohyponyms)	MRR	MAP
base sample, 0%	39.31	25.77
cohyponyms, 25%	39.54±0.51	24.95±0.69
cohyponyms, 50%	39.37±0.6	24.91±0.41
cohyponyms, 75%	38.01±0.55	23.86±0.5
cohyponyms, 100%	38.93	24.54
non-cohyponyms, 25%	39.49±0.38	25.19±0.48
non-cohyponyms, 50%	40.17±0.42	25.7±0.24
non-cohyponyms, 75%	38.74±0.44	24.77±0.55
non-cohyponyms, 100%	**41.55**	**26.29**
random, 25%	39.14±0.53	24.91±0.34
random, 50%	39.94±1.24	25.32±0.77
random, 75%	39.95±0.8	25.27±0.25
random, 100%	39.72±0.27	25.34±0.12

Table 7. WordNet training samples by type and proportion of removed nodes with metrics achieved by corresponding fine-tuned models.

at 25% proportion. These findings suggest that cohyponyms play a crucial role in the model's learning process, as their reduction corresponds to decreased performance. Furthermore, increasing the concentration of cohyponyms by selectively removing items with other taxonomic relations to test items resulted in improved performance metrics.

Fig. 4 reveals a non-monotonic relationship between the proportion of removed items and performance scores in both cohyponym and non-cohyponym conditions. The scores demonstrate an initial increase from 0 to 50% removal, followed by a decrease reaching their minimum at 75% removal. This pattern suggests the absence of a direct correlation between sample size and model performance. This observation aligns with our findings in Section 6.1, where we examined the performance of base models trained on datasets with removed test nodes from different origins.

Analysis of random samples reveals no significant variation between values. The scores stabilize at levels comparable to those observed at 25-50% removal proportions in the other two experimental conditions. This pattern can be attributed to the relatively low baseline frequency of cohyponyms

in the sample, which ranges from approximately 4% to 15% across all proportions. As previously demonstrated, the systematic bias introduced by selective removal of non-cohyponyms while retaining cohyponyms becomes particularly pronounced when the removal proportion exceeds 50%.

These findings suggest that the sampling methodology influences the observed patterns in the data and provide support for our hypothesis regarding the bias introduced by learning hypernyms for the cohyponyms of the test items.

6.3. Fine-tuning on WordNet with definitions. The metrics obtained after training with definitions included in the prompt for certain items are presented in Tab. 8 (inclusion by node type, the first setting) and 9 (omission by node type, the second setting), and the MRR scores are depicted in Fig. 5 and 6, respectively.

percentage of items with definition by number of cohyponyms	MRR	MAP
base sample, +0%	39.31	**25.77**
cohyponyms, +25%	39.94	25.73
cohyponyms, +50%	39.2	25.01
cohyponyms, +75%	39.45	24.57
cohyponyms, +100%	36.59	23.18
non-cohyponyms, +25%	**40.31**	25.23
non-cohyponyms, +50%	39.35	24.72
non-cohyponyms, +75%	39.3	24.42
non-cohyponyms, +100%	38.43	23.54
random, +25%	39.68	25.06
random, +50%	39.19	24.4
random, +75%	39.49	24.45
random, +100%	38.45	23.41

Table 8. WordNet training samples by proportion of added definitions and type of nodes to which they were added with metrics achieved by corresponding fine-tuned models (1st definition setting)

In the first experimental setting, differences are apparent only at the lowest (25%) and highest (100%) proportions. In both cases, adding definitions to cohyponyms results in worse performance compared to adding

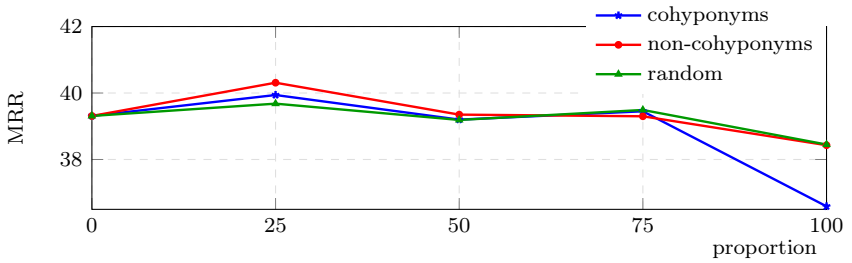

Figure 5. MRR across three conditions in the 1st definition setting.

percentage of items with definition by number of cohyponyms	MRR	MAP
base sample, 100%	36.66	21.20
cohyponyms, -25%	38.41	22.15
cohyponyms, -50%	39.76	**22.93**
cohyponyms, -75%	**39.85**	22.73
cohyponyms, -100%	39.73	22.39
non-cohyponyms, -25%	36.82	21.39
non-cohyponyms, -50%	36.74	21.45
non-cohyponyms, -75%	35.81	20.61
non-cohyponyms, -100%	36.41	22.07
random, -25%	36.94	21.59
random, -50%	37.10	21.88
random, -75%	36.96	21.8
random, -100%	37.22	21.86

Table 9. WordNet training samples by proportion of nodes from which we excluded definitions and type of such nodes with metrics achieved by corresponding fine-tuned models (2nd definition setting).

definitions to non-cohyponyms, especially at the highest proportion, where we observe a sharp decrease in performance. One possible explanation is that, in this setup, all sister terms were learned in a form that differed significantly from the format used during the extra fine-tuning and evaluation

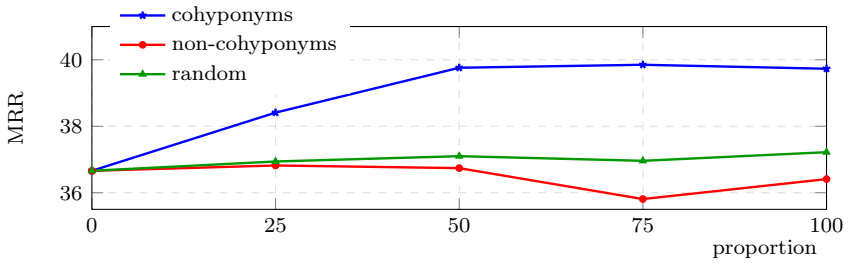

Figure 6. MRR across three conditions in the 2nd definition setting.

phases, as the definitions were not provided for the "1A: English" dataset. As a result, the model is unable to effectively retrieve this information from memory.

In the second setting, we observe that the absence of cohyponym definitions led to superior performance compared to other conditions across all tested proportions. This finding contradicts the general assumption proposed by Moskvoretskii et al. [25] regarding the benefits of including definitions during initial training. However, it supports our hypothesis. Specifically, since cohyponyms contribute more significantly than other items, performance improves when the prompt structure more closely matches the one used during evaluation. This is precisely the case for "1A: English", where the datasets did not include definitions, and thus the evaluation was conducted using prompts without definitions.

A comparison of the two settings reveals that, for our dataset, the impact of definitions appears to contradict the claims made by the authors of TaxoLLaMA. In the first setting, the best performance was achieved by models with the lowest percentage of added definitions, while in the second setting, the highest performance corresponded to models with the higher percentage of removed definitions. Furthermore, models in the second setting generally performed worse than those in the first setting. The average metrics for the second setting were MRR 39.13 and MAP 24.58, compared to the first setting, where the average metric scores were MRR 37.57 and MAP 21.85.

In this subsection, we further investigated the impact of node types based on their taxonomic relationships to the test items. While the bias

observed in the first section was minor, in the second setting we demonstrated that the performance of the model improves when definitions are omitted from the cohyponyms. This indicates that a a consistent form of presentation of cohyponyms during training and the test items during evaluation is critical.

6.4. Fine-tuning on "1A: English" training dataset. In Tab. 10, 11 and 12 we compare metric scores that were achieved by the models trained in subsection 6.2 and 6.3 after additional fine-tuning. Fig. 7, 8, and 9 illustrate the improvements in MRR scores attained by the model across all conditions in our three experimental settings.

percentage of deleted nodes (*by number of cohyponyms*)	*MRR*(*+gain*)	*MAP*(*+gain*)
base sample, 100%	$\mathbf{52.21}_{+12.9}$	$\mathbf{33.51}_{+7.74}$
cohyponyms, 25%	$51.47_{+11.93}$	$33.35_{+8.4}$
cohyponyms, 50%	$50.75_{+11.38}$	$32.83_{+7.92}$
cohyponyms, 75%	$51.18_{+13.18}$	$33.18_{+9.32}$
cohyponyms, 100%	$49.72_{+10.79}$	$32.31_{+7.77}$
non-cohyponyms, 25%	$50.9_{+11.41}$	$32.9_{+7.71}$
non-cohyponyms, 50%	$51.56_{+11.39}$	$33.25_{+7.55}$
non-cohyponyms, 75%	$50.78_{+12.04}$	$32.85_{+8.08}$
non-cohyponyms, 100%	$51.54_{+9.99}$	$33_{+6.71}$
random, 25%	$51.32_{+12.19}$	$33.16_{+8.25}$
random, 50%	$51.58_{+11.64}$	$33.27_{+7.95}$
random, 75%	$50.22_{+10.27}$	$32.78_{+7.51}$
random, 100%	$51.47_{+11.75}$	$33.3_{+7.96}$

Table 10. Metrics obtained by the models discussed in subsection 6.2 after fine-tuning on the "1A: English" training dataset. The gain reflects the difference between the second and the first stage of fine-tuning, detailed in Tab. 7, where types and proportions of removed nodes in each sample is determined

Based on the results presented in the tables, we observe that additional fine-tuning improves overall performance but also introduces slight changes in the inter-group patterns previously identified in subsection 6.2. The distribution of maximum metric values differs between the two procedures. In

percentage of nodes with definitions (by number of cohyponyms)	*MRR(+gain)*	*MAP(+gain)*
base sample, 0%	**52.21**$_{+12.9}$	33.51$_{+7.74}$
cohyponyms, +25%	51.79$_{+11.85}$	33.35$_{+7.62}$
cohyponyms, +50%	51.3$_{+12.1}$	33.17$_{+8.16}$
cohyponyms, +75%	51.51$_{+12.06}$	33.31$_{+8.74}$
cohyponyms, +100%	50.28$_{+13.69}$	32.49$_{+9.31}$
non-cohyponyms, +25%	51.48$_{+11.17}$	33.09$_{+7.86}$
non-cohyponyms, +50%	51.2$_{+11.85}$	33.27$_{+8.55}$
non-cohyponyms, +75%	51.66$_{+12.36}$	33.44$_{+9.02}$
non-cohyponyms, +100%	51.38$_{+12.95}$	33.18$_{+9.64}$
random, +25%	51.79$_{+12.11}$	**33.62**$_{+8.56}$
random, +50%	51.42$_{+12.23}$	33.2$_{+8.8}$
random, +75%	51.56$_{+12.07}$	33.39$_{+8.94}$
random, +100%	50.1$_{+11.65}$	32.71$_{+9.3}$

Table 11. Metrics, achieved by models discussed in subsection 6.3 (1st setting) after fine-tuning on "1A: English" training dataset. The gain reflects the difference between the second and the first stage of fine-tuning, detailed in Tab. 8, where node types and proportions of added definitions in each sample is determined

the non-cohyponym removal condition, the new maximum was achieved on the sample that had previously ranked second-highest (cf. Table 7 and 10).

A similar shift is observed in the first definition setting, where the top result in the non-cohyponym condition changed from 25% to 75% (cf. Tab 8 and 11). In the second definition setting, all maximum scores (except one MAP score in the non-cohyponym condition) were obtained with 100% of the definitions removed (Tab. 12). This pattern is consistent with our previous observations that cohyponyms significantly affect performance and that providing definitions for too many items has a negative impact. Fine-tuning on the "1A: English" dataset was performed without definitions, so the samples with the maximum number of nodes with definitions removed appear to align better with the secondary fine-tuning, resulting in higher metric scores.

Upon closer examination of the secondary fine-tuning gains, we find that the metrics achieved after the secondary fine-tuning are comparable across

percentage of nodes with removed definitions (by number of cohyponyms)	*MRR(+gain)*	*MAP(+gain)*
base sample, 100%	$49.25_{+12.59}$	$32.04_{+10.84}$
cohyponyms, -25%	$50.2_{+11.79}$	$32.6_{+10.45}$
cohyponyms, -50%	$50.56_{+10.8}$	32.93_{+10}
cohyponyms, -75%	$50.23_{+10.38}$	$32.23_{+9.5}$
cohyponyms, -100%	$51.21_{+11.48}$	$33.34_{+10.95}$
non-cohyponyms, -25%	$51.22_{+14.4}$	$33.34_{+11.95}$
non-cohyponyms, -50%	$49.89_{+13.15}$	$32.77_{+11.32}$
non-cohyponyms, -75%	$50.69_{+14.88}$	$32.95_{+12.34}$
non-cohyponyms, -100%	$51.29_{+14.88}$	$33.14_{+11.07}$
random, -25%	$50.34_{+13.4}$	$33.15_{+11.56}$
random, -50%	$50.87_{+13.77}$	$33.02_{+11.14}$
random, -75%	$51.58_{+14.62}$	$33.19_{+11.39}$
random, -100%	$\mathbf{51.7}_{+14.48}$	$\mathbf{33.86}_{+12}$

Table 12. Metrics, achieved by models discussed in subsection 6.3 (2nd setting) after fine-tuning on "1A: English" training dataset. The gain reflects the difference between the second and the first stage of fine-tuning, detailed in Tab. 9, where node types and proportions of removed definitions in each sample is determined

all three experimental conditions. Specifically, in the removal experiment, the comparison of the results after the two stages reveals that performance improvements in the non-cohyponym condition are less pronounced than in the cohyponym condition (see Fig. 9).

This pattern extends to the definition settings, where as well there is an inverse relationship between performance after initial fine-tuning and the subsequent gains from the second stage. This is supported by the negative correlation measured by Pearson's r, which is -0.8875 for MRR and -0.9669 for MAP.

These findings suggest that fine-tuning offers only a limited performance improvement, regardless of the presence of cohyponym bias.

However, it is important to note that the "1A: English" dataset itself contains a small proportion of cohyponyms—approximately 3%. Therefore,

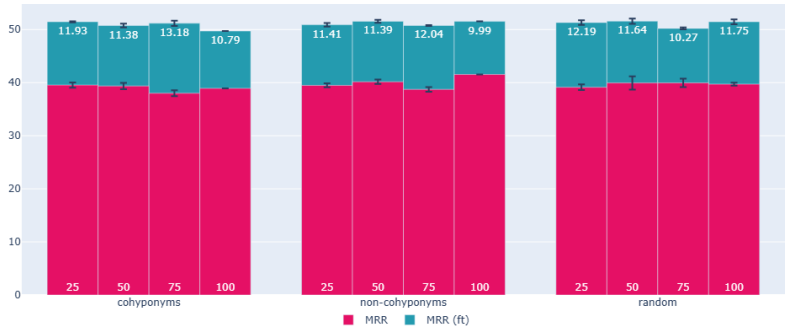

Figure 7. The gain in MRR the models initially fine-tuned on WordNet (removal setting) achieved after the secondary fine-tuning on "1A: English" dataset.

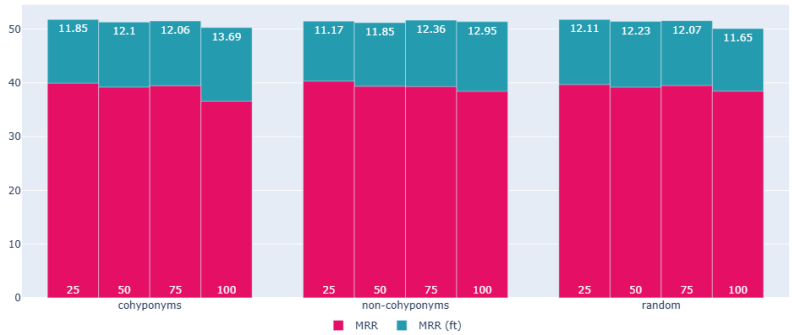

Figure 8. The gain in MRR the models initially fine-tuned on WordNet (1st definition setting) achieved after the secondary fine-tuning on "1A: English" dataset.

it is plausible that the cohyponym bias observed during secondary fine-tuning was already present after the initial fine-tuning. The differences we observe may thus result from the overlap of these two biases. Nonetheless, this remains a hypothesis that we have not been able to verify.

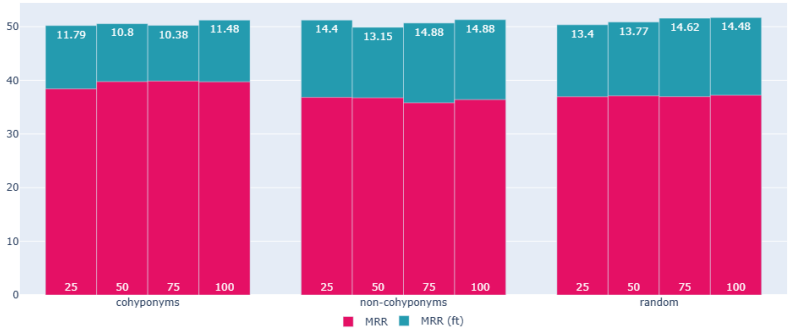

Figure 9. The gain in MRR the models initially fine-tuned on WordNet (2nd definition setting) achieved after the secondary fine-tuning on "1A: English" dataset.

§7. Conclusion

Our study highlights the significant impact of cohyponyms on the performance of model in the Hypernym Discovery task, reinforcing our argument that their presence constitutes a form of data leakage. This phenomenon is attributable to the similar contexts in which cohyponyms occur, resulting in highly similar embeddings. By showing that cohyponyms serve as implicit cues for identifying hypernyms, our results emphasize the necessity of excluding them during training to achieve unbiased evaluations that more accurately reflect real-world scenarios. The variability in performance when cohyponyms are removed, contrasted with the minimal impact of other manipulations like random deletions, underscores the nuanced relationship between the presence of cohyponyms and task outcomes. These insights highlight the importance of meticulous dataset curation to ensure the reliability and applicability of taxonomy expansion systems in practical contexts where cohyponyms are not necessarily present.

Moreover, our comparative analysis indicates that fine-tuning and experimental adjustments introduce complexities that can unpredictably alter performance trends. The observed shifts in metrics, influenced by cohyponym density and dataset characteristics such as the presence of definitions, challenge prior assumptions and suggest that evaluation frameworks must account for these intricacies. Collectively, our findings advocate for

a more rigorous approach to sample design and stress the importance of mitigating cohyponym bias in training data.

A potential solution, not addressed in [25], is the use of diachronic WordNet datasets as proposed in [27]. These datasets involve training on nodes from an older version of WordNet and testing on a newer version, leveraging naturally acquired collections due to chronological continuity. This method appears unbiased and better mirrors the practical conditions of the taxonomy enrichment process.

Acknowledgments

The computations were performed utilizing the Yandex Datasphere service, with support from the Non-commercial Foundation for Support of Science and Education "INTELLECT".

References

1. A. I. A. Aldine, M. Harzallah, G. Berio, N. Bechet, and A. Faour, *Redefining Hearst Patterns by Using Dependency Relations*, in: 10th International Conference on Knowledge Engineering and Ontology Development, 2018, pp. 148–155.
2. H. Babaei Giglou, J. D'Souza, and S. Auer, *LLMs4OL: Large Language Models for Ontology Learning*, in: International Semantic Web Conference, Cham: Springer Nature Switzerland, 2023, pp. 408–427.
3. Y. Bai, R. Zhang, F. Kong, J. Chen, and Y. Mao, *Hypernym Discovery via a Recurrent Mapping Model*, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP, 2021, pp. 2912–2921.
4. M. Baroni and A. Lenci, *How We Blessed Distributional Semantic Evaluation*, in: Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics, 2011, pp. 1–10.
5. M. Baroni, N.-Q. Do, and C.-c. Shan, *Entailment Above the Word Level in Distributional Semantics*, in: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 23–32.
6. G. Berend, M. Makrai, and P. Foldiak, *300-sparsans at SemEval-2018 Task 9: Hypernymy as Interaction of Sparse Attributes*, in: Proceedings of the 12th International Workshop on Semantic Evaluation, 2018, pp. 928–934.
7. G. Bernier-Colborne and C. Barriere, *CRIM at SemEval-2018 Task 9: A Hybrid Approach to Hypernym Discovery*, in: Proceedings of the 12th International Workshop on Semantic Evaluation, 2018, pp. 725–731.
8. S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, Inc., 2009.
9. G. Bordea, E. Lefever, and P. Buitelaar, *SemEval-2016 Task 13: Taxonomy Extraction Evaluation (TexEval-2)*, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016.

10. J. Camacho-Collados, C. Delli Bovi, L. Espinosa Anke, S. Oramas, T. Pasini, E. Santus, V. Shwartz, R. Navigli, and H. Saggion, *SemEval-2018 Task 9: Hypernym Discovery*, in: Proceedings of the 12th International Workshop on Semantic Evaluation, 2018, pp. 712–724.
11. C. Chen, K. Lin, and D. Klein, *Constructing Taxonomies from Pretrained Language Models*, Computing Research Repository, arXiv:2010.12813 (2020), pp. 4687–4700.
12. K. Erk, *Vector Space Models of Word Meaning and Phrase Meaning: A Survey.* — Language and Linguistics Compass **6**.10 (2012), 635–653.
13. H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, *et al.*, *Scaling Instruction-Finetuned Language Models.* — Journal of Machine Learning Research **25**(70) (2024), 1–53.
14. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*, in: J. Burstein, C. Doran, and T. Solorio (eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.
15. M. Hanna and D. Mareček, *Analyzing BERT's Knowledge of Hypernymy via Prompting*, in: Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, 2021, pp. 275–282.
16. M. A. Hearst, *Automatic Acquisition of Hyponyms from Large Text Corpora*, in: COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics, 1992, pp. 539–545.
17. W. Held and N. Habash, *The Effectiveness of Simple Hybrid Systems for Hypernym Discovery*, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 3362–3367.
18. M. Jiang, X. Song, J. Zhang, and J. Han, *TaxoEnrich: Self-Supervised Taxonomy Completion via Structure-Semantic Representations*, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 925–934.
19. A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, *et al.*, *Mistral 7B*, Computing Research Repository, arXiv:2310.06825 (2023).
20. D. Jurgens and M. T. Pilehvar, *SemEval-2016 Task 14: Semantic Taxonomy Enrichment*, in: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), 2016, pp. 1092–1102.
21. J. Liao, X. Chen, and L. Du, *Concept Understanding in Large Language Models: An Empirical Study*, Tiny Paper (ICLR), 2023, pp. 1–5.
22. Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, Computing Research Repository, arXiv:1907.11692 (2019).
23. G. A. Miller, *WordNet: A Lexical Database for English.* — Communications of the ACM **38**(11) (1995), 39–41.
24. V. Moskvoretskii, E. Neminova, A. Lobanova, A. Panchenko, and I. Nikishina, *TaxoLLaMA: WordNet-Based Model for Solving Multiple Lexical Sematic Tasks*, Computing Research Repository, arXiv:2403.09207 (2024).

25. V. Moskvoretskii, E. Neminova, A. Lobanova, A. Panchenko, and I. Nikishina, *Large Language Models for Creation, Enrichment and Evaluation of Taxonomic Graphs*. — Semantic Web, forthcoming.
26. N. Nakashole, G. Weikum, and F. Suchanek, *PATTY: A Taxonomy of Relational Patterns with Semantic Types*, in: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012, pp. 1135–1145.
27. I. Nikishina, V. Logacheva, A. Panchenko, and N. Loukachevitch, *RUSSE'2020: Findings of the First Taxonomy Enrichment Task for the Russian language*, Computing Research Repository, arXiv:2005.11176 (2020).
28. I. Nikishina, M. Tikhomirov, V. Logacheva, Y. Nazarov, A. Panchenko, and N. Loukachevitch, *Taxonomy Enrichment with Text and Graph Vector Representations*. — Semantic Web **13**(3) (2020), 441–475.
29. I. Nikishina, P. Chernomorchenko, A. Demidova, A. Panchenko, and C. Biemann, *Predicting Terms in Is-A Relations with Pre-Trained Transformers*, in: Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings), 2023, pp. 134–148.
30. A. Ravichander, E. Hovy, K. Suleman, A. Trischler, and J. C. K. Cheung, *On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT*, in: Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics, 2020, pp. 88–102.
31. A. Ritter, S. Soderland, and O. Etzioni, *What Is This, Anyway: Automatic Hypernym Discovery*, in: AAAI Spring Symposium: Learning by Reading and Learning to Read, 2009, pp. 88–93.
32. S. Roller, K. Erk, and G. Boleda, *Inclusive Yet Selective: Supervised Distributional Hypernymy Detection*, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 1025–1036.
33. S. Roller, D. Kiela, and M. Nickel, *Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora*, Computing Research Repository, arXiv:1806.03191 (2018).
34. K. Sabirova and A. Lukanin, *Automatic Extraction of Hypernyms and Hyponyms from Russian Texts*, in: AIST (supplement), 2014, pp. 35–40.
35. J. Seitner, C. Bizer, K. Eckert, S. Faralli, R. Meusel, H. Paulheim, and S. Ponzetto, *A Large Database of Hypernymy Relations Extracted from the Web*, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 360–367.
36. J. Shen, Z. Shen, C. Xiong, C. Wang, K. Wang, and J. Han, *TaxoExpan: Self-Supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network*, in: Proceedings of the Web Conference 2020, 2020, pp. 486–497.
37. A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-j. (Paul) Hsu, and K. Wang, *An Overview of Microsoft Academic Service (MAS) and Applications*, in: Proceedings of the 24th International Conference on World Wide Web, 2015.
38. R. Snow, D. Jurafsky, and A. Ng, *Learning Syntactic Patterns for Automatic Hypernym Discovery*, in: Advances in Neural Information Processing Systems **17**, 2004, pp. 1297–1304.

39. M. Tikhomirov and N. Loukachevitch, *Meta-Embeddings in Taxonomy Enrichment Task*, in: Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference "Dialogue", 2021.
40. M. Tikhomirov and N. Loukachevitch, *Exploring Prompt-Based Methods for Zero-Shot Hypernym Prediction with Large Language Models*, Computing Research Repository, arXiv:2401.04515 (2024).
41. H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, *Llama 2: Open Foundation and Fine-Tuned Chat Models*, Computing Research Repository, arXiv:2307.09288 (2023).
42. D. Ustalov, N. Arefyev, C. Biemann, and A. Panchenko, *Negative Sampling Improves Hypernymy Extraction Based on Projection Learning*, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, 2017, Vol. 2, pp. 543–551.
43. J. Weeds, D. Clarke, J. Reffin, D. Weir, and B. Kelle, *Learning to Distinguish Hypernyms and Co-Hyponyms*, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 2249–2259.
44. H. Xu, Y. Chen, Z. Liu, Y. Wen, and X. Yuan, *TaxoPrompt: A Prompt-Based Generation Method with Taxonomic Context for Self-Supervised Taxonomy Expansion*, in: IJCAI, Vol. 22, 2022, pp. 4432–4438.
45. J. Yamane, T. Takatani, H. Yamada, M. Miwa, and Y. Sasaki, *Distributional Hypernym Generation by Jointly Learning Clusters and Projections*, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 1871–1889.
46. G. Yun, Y. Lee, A.-S. Moon, and J. Lee, *Hypert: Hypernymy-Aware BERT with Hearst Pattern Exploitation for Hypernym Discovery.* — Journal of Big Data **10**(1) (2023), 141.
47. J. Zhang, X. Song, Y. Zeng, J. Chen, J. Shen, Y. Mao, and L. Li, *Taxonomy Completion via Triplet Matching Network*, in: Proceedings of the AAAI Conference on Artificial Intelligence, **35**, 2021, pp. 4662–4670.

RAS Institute of Linguistics,
Lomonosov Moscow State University
E-mail: sadkovsky@iling-ran.ru

Lomonosov Moscow State University
E-mail: louk_nat@mail.ru
E-mail: igrishin@sev.msu.ru

Поступило 28 февраля 2025 г.