TaxoRankConstruct: A Novel Rank-based Iterative Approach to Taxonomy Construction with Large Language Models

Oleksandr Marchenko[1, 2] and Danylo Dvoichenkov[1, *]

[1] International Research and Training Center for Information Technologies and Systems, 40, Akademika Glushkova Ave, Kyiv, 03187, Ukraine
[2] Taras Shevchenko National University of Kyiv, 60, Volodymyrska St, Kyiv, 01033, Ukraine

Abstract

Paper presents a novel method for the construction of taxonomical classifications (concept hierarchies) for concepts using large language models. Traditional methods of taxonomy construction often focus heavily on hypernym-hyponym relationships, emphasizing hierarchical connections between concepts. However, these approaches tend to overlook the qualitative attributes of objects that form the foundation of classification. In contrast, the approach proposed in this paper is based on the premise that "the properties of objects are primary, while the types of objects are secondary." This foundational idea drives the development of TaxoRankConstruct, a novel rank-based iterative approach that leverages Large Language Models (LLMs) to construct more nuanced taxonomies. This method aims to enhance the clarity and precision of taxonomical hierarchies by systematically organizing concepts based on specific, identifiable characteristics.

Keywords

Taxonomy Construction, Large Language Models, Hierarchical Classification, Ontology Learning, Concept Hierarchies, Natural Language Processing, Human-AI Collaboration, Iterative Methods

1. Introduction

Taxonomies are essential tools across various disciplines, facilitating the organization of knowledge by classifying concepts based on shared characteristics [1]. They are widely used in fields like biology, information science, astronomy, and chemistry, providing a structured framework for managing data and concepts [2]. However, constructing large-scale taxonomies from scratch remains a significant challenge, particularly when predefined hierarchies are unavailable, and dynamic criteria must be considered for iterative exploration and refinement.

Despite advancements in natural language processing, there is still a gap in methodologies capable of building comprehensive taxonomies from scratch. Traditional approaches depend on expert-driven categorizations or clustering techniques to organize existing concepts into hierarchical structures. These methods often rely on predefined similarities and known concepts, limiting their capacity to iteratively explore and refine taxonomies using dynamic criteria [3, 4, 5]. The lack of automated or semi-automated tools that can adapt to new data and build taxonomies from the ground up underscores the need for more flexible and innovative solutions [6].

Our proposed method, TaxoRankConstruct, addresses this gap by introducing a rank-based iterative approach to building taxonomies from scratch. It identifies a set of "taxonomical ranks" for

Information Technology and Implementation (IT&I-2024), November 20-21, 2024, Kyiv, Ukraine

[*] Corresponding author.

supersokol777@gmail.com (D. Dvoichenkov); rozenkrans17@gmail.com (O.Marchenko);
0009-0007-1935-6743 (D. Dvoichenkov); 0000-0002-5408-5279 (O.Marchenko);

the root concept, using these predefined ranks to determine the taxonomy's depth and the criteria by which concepts differ. This method enables the systematic and transparent exploration of concepts based on their taxonomical properties, supporting iterative population of taxonomies with selected criteria and offering a flexible solution for taxonomy construction [7, 8].

The primary goal of this research is to develop a new method for iterative taxonomy construction, emphasizing the multiple ways a single concept can be classified. This raises important questions about the best approach to algorithmic, iterative taxonomy creation: Should we focus on examining concepts and their properties sequentially, or should we first explore the root concept to identify the general properties that shape the entire taxonomy?

By introducing taxonomical ranks and criteria, our method enhances the ability to generate and evaluate taxonomies more effectively. We also present human evaluation results and statistics on the generated taxonomical classifications. Additionally, we address some of the shortcomings of existing taxonomies, providing insights into how our approach complements and improves current practices. Our results are designed to be reproducible, and the proposed approach is highly adaptable, making it suitable for application across many domains. This flexibility allows for adjustment to meet specific needs and contexts.

In the remainder of this paper, we explore the proposed methodology and its evaluation in detail. Section 1 reviews the related work. Section 2 establishes the conceptual framework that underpins our approach, defining the key concepts and challenges associated with effective taxonomy construction using large language models (LLMs). Section 3 outlines the detailed methodology employed in this study. Section 4 describes the experimental setup and scenarios designed to rigorously test our approach, while Section 5 discusses the evaluation and results, focusing on human assessment of the quality and relevance of the generated taxonomies. Finally, Section 6 explores potential applications of our work, suggests directions for future research, and concludes with a summary of the study's key contributions.

2. Related Work

The construction of taxonomical classifications has been extensively researched across various disciplines, as previously mentioned. This section reviews recent advancements and methodologies in taxonomy construction, highlighting their strengths and limitations.

2.1. Supervised and Semi-Supervised Methods

Traditional taxonomy construction methods often use supervised and semi-supervised learning techniques. These methods typically extract lexical features and train classifiers to identify hypernym-hyponym relationships from curated datasets. For example, the methods proposed by Fu et al. [9] use word embeddings to classify relations between terms, while order embedding techniques represent partial orders between words [10]. However, these approaches are limited by the availability of annotated data and their adaptability to domain-specific texts.

2.2. Unsupervised Methods

Unsupervised methods aim to build taxonomies without relying on labeled data. For instance, TaxoGen employs adaptive term embedding and clustering to create topic taxonomies in a top-down manner. This approach uses term embeddings to recursively split topics into finer subtopics, addressing challenges related to semantic granularity and coherence at different taxonomy levels [3]. Another notable unsupervised method is CoRel, which uses seed-guided learning to expand a tree-structured seed taxonomy provided by users. CoRel's relation transferring module helps discover new topics and subtopics by capturing relationships between terms in the corpus [4]. A recent study by Mishra et al. introduced the FLAME (Self-Supervised Low-Resource Taxonomy Expansion using Large Language Models) method, which focuses on expanding taxonomies in low-resource environments. By leveraging a self-supervised approach, FLAME proves effective in scenarios where minimal labeled data is available for high-quality taxonomy generation. This method provides an important solution for taxonomy expansion tasks, particularly in cases where traditional methods require extensive resources [11]. TaxoClass offers a novel approach for hierarchical multi-label text classification using only class names. It simulates human experts by identifying core classes for each document and then generalizes the classifier through multi-label self-training, significantly improving performance over previous methods [12]. TaxoCom applies hierarchical discovery of novel topic clusters to complete a user-provided partial hierarchy by recursively expanding it with new topics and subtopics [5]. WERECE uses word embedding refinement for educational concept extraction, integrating manifold learning and semantic clustering to adapt pre-trained models for subject-specific concepts, achieving high precision and recall [13].

2.3. Use of Large Language Models (LLMs)

The rise of large language models (LLMs) has significantly impacted taxonomy construction methodologies. LLMs like GPT-3 and BERT have been used in both prompting and fine-tuning paradigms to generate taxonomies. A comparative study by researchers highlighted the effectiveness of few-shot prompting, where a few examples guide the LLM in generating the desired taxonomy structure. This approach is useful for generating taxonomies from limited data but may struggle with less powerful models [7]. Another method, Chain-of-Layer (CoL), proposes an iterative prompting technique where LLMs build taxonomies layer by layer. This method ensures that the taxonomy follows hierarchical constraints and reduces issues like hallucination and incorrect parent-child relations by using an ensemble-based ranking filter [8]. The Hierarchical Prompting Taxonomy (HPT) uses five different prompting strategies: Role Prompting, Zero-Shot Chain-of-Thought Prompting (Zero-CoT), Three-Shot Chain-of-Thought Prompting (3-CoT), Least-to-Most Prompting, and Generated Knowledge Prompting (GKP). This method allows LLMs to systematically address tasks by moving from simple to complex prompts, enhancing the model's problem-solving capabilities [14]. Another innovative method involves iterative prompting with frequency analysis to refine taxonomy construction. This technique uses frequent token analysis to improve the accuracy and completeness of the generated taxonomies, addressing issues like domain shifts and attribute inflation [15]. The Human-AI Collaborative Taxonomy Construction method combines human expertise with AI-generated concepts. Here, LLMs produce initial taxonomy structures that are then reviewed and refined by human experts. This collaborative approach improves the quality and accuracy of the final taxonomies [16]. Additionally, the Modular Ontology Modeling (MOMo) approach facilitates ontology construction by creating compact, independent modules. These modules encapsulate key concepts and their main features, streamlining maintenance and enhancing flexibility and adaptability [17]. Ontology-Enhanced Representation Learning integrates ontological knowledge into embedding models through contrastive learning. This method generates synthetic concept definitions and creates semantically related text pairs by synonym substitution, improving the model's understanding of ontological relations [18]. The LLMs4OL (Large Language Models for Ontology Learning) paradigm provides a comprehensive framework for automated ontology construction. This approach involves tasks such as term typing, type taxonomy discovery, and non-taxonomic relationship extraction. Each task leverages LLMs to accelerate ontology learning, using datasets like GeoNames and Schema.Org [19]. Finally, Ontology Engineering with LLMs uses prompt engineering to transform natural language statements into formal logical expressions suitable for ontology description languages like OWL. This involves advanced prompting techniques and fine-tuning strategies to enhance the model's performance in formalizing ontological statements [6].

2.4. Challenges and Limitations

Despite advancements in taxonomy construction using LLMs, several significant challenges remain. One major issue is the tendency of LLMs to hallucinate, generating incorrect or irrelevant relations that compromise taxonomy quality. Tools like CoL attempt to mitigate this problem by filtering out invalid relations, but further improvements are needed to enhance system reliability [8]. Additionally, while supervised and semi-supervised methods offer precise control over taxonomy construction, they heavily depend on extensive labeled data, which is not always feasible, especially in domain-specific applications [4]. Furthermore, existing tools like CoRel and TaxoGen have limitations in generating taxonomies from scratch. For example, CoRel uses seed-guided learning to expand pre-existing taxonomies but lacks a framework for building entirely new taxonomies based on newly identified concepts and their properties [4]. Similarly, TaxoGen relies on clustering techniques but does not provide the flexibility to define taxonomical ranks that can adapt to evolving datasets and domains [3]. Moreover, these methods do not support the iterative enrichment of taxonomies by dynamically adjusting to different classification criteria, highlighting the need for more advanced approaches that can construct and refine taxonomies to accommodate the dynamic nature of data and emerging concepts.

3. Conceptual Framework

In this section, we will establish the foundational concepts and terminology essential for understanding the taxonomy construction method proposed in this research. This foundational framework is crucial for understanding the subsequent "Methodology" section, where we will detail the practical steps involved in constructing taxonomy.

3.1. Taxonomy as a Tree of Concepts

For the purposes of this study, we consider taxonomy *T* as a tree composed of a set of concepts, denoted as *C*.

Although taxonomies can have more complex structures, such as graphs with multiple interconnections, we simplify our analysis by assuming a strictly hierarchical tree structure. This simplification allows for a more straightforward approach to organizing and analyzing concepts within the taxonomy.

3.2. Subconcept Formation Based on Object Properties

Within the given taxonomy *T*, a concept C_i has a set of subconcepts M_i if and only if all objects classified under concept C_i share a specific set of properties *F*, where $|F| > 1.$ Among these properties, $|F| - 1$ are consistent across all subconcepts in M_i, while a single property F_j can vary, leading *to* $V = |M_i|$ different values, and there is a bijection $f: F_j{}^{\leftrightarrow}M_i$

This approach ensures that the classification is grounded in the inherent attributes of the objects rather than arbitrary hierarchical relationships.

3.3. Uniform Property Distribution across Concepts

All concepts within a given taxonomy *T* possess a consistent set of properties *F*. This means that the parent concept inherently includes all potential properties of its subconcepts, although some of these properties may remain undefined or unknown. For example, the concept '*spoon*' shares the property '*material*' with the concept '*iron spoon*'; however, while '*material*' is defined as '*iron*' for the '*iron spoon*,' it may be undefined or '*unknown*' for the broader concept '*spoon*.' Additionally, some properties might have the value '*absent,*' such as the property '*presence of a notochord*' in concepts like '*prokaryotes*' or '*fungi.*'

3.4. Taxonomy Depth and Property Count

The depth of the taxonomy T is determined by the number of properties F that its concepts possess:

$$d(T) = |F|,$$

This approach to defining depth allows for a more meaningful metric in understanding the complexity of the taxonomy, as it directly correlates with the diversity of attributes represented within the hierarchical structure.

3.5. Multitaxonomy Approach

While the initial assumption was to consider the taxonomy as a single tree structure, further analysis led to a more sophisticated approach: the use of a set of trees, where each tree corresponds to a different set of properties. This resulted in the concept of "multitaxonomies," where each taxonomy consists of multiple trees of varying depths. For instance, in the context of "Device," one tree might represent the hierarchy "*Operating System > Device Type > Form Factor,*" while another might represent "*Manufacturer > Model > Series.*"

This multitree approach allows for a more nuanced representation of concepts, accommodating different perspectives and categorizations within the same domain. Although a fully developed taxonomy should ideally integrate all concepts into a single complex graph, this study focuses on this intermediate step of multitaxonomies. This approach serves as a bridge between traditional single-tree taxonomies and more advanced graph-based structures, which will be explored in future work. By leveraging multiple trees, we can capture the diversity of object classifications without forcing all properties into a single hierarchical structure.

3.6. Finalizing Key Concepts

In our discussion so far, we have introduced the concepts of taxonomy T, the set of concepts C, individual concepts C_j, the set of subconcepts M_i, and the set of properties F.

Now, we introduce a specific concept R, known as the Root Concept. With the introduction of the notion of multitaxonomy, we redefine T to represent a collection of taxonomies, denoted as T_i. This means that T is no longer a single taxonomy, but rather a set of taxonomies with elements T_i, each associated with its own set of properties F_i. Consequently, F now represents a set of property sets, encompassing all the individuals F_i associated with each taxonomy T_i.

For each taxonomy T_i, the corresponding set of concepts is denoted as C_i, and within each C_i, an individual concept is represented as C_{ij} (where $0 < \mathrm{i} \leq |\mathrm{T}|$ and $0 < \mathrm{j} \leq |C_i|$). Similarly, the set of subconcepts within C_{ij} is represented as M_{ij}.

Figure 1 provides a clear example of a root concept, taxonomical ranks, and sub-concepts. It visually demonstrates how these elements are structured in a multitaxonomy framework.

Having established the key concepts and the framework for our approach, we are now prepared to explain the specifics of how the proposed method operates. In the next section, "Methodology," we will explore the practical application of this framework, detailing the step-by-step process for constructing a multitaxonomy and identifying the full set of subconcepts and their relationships.

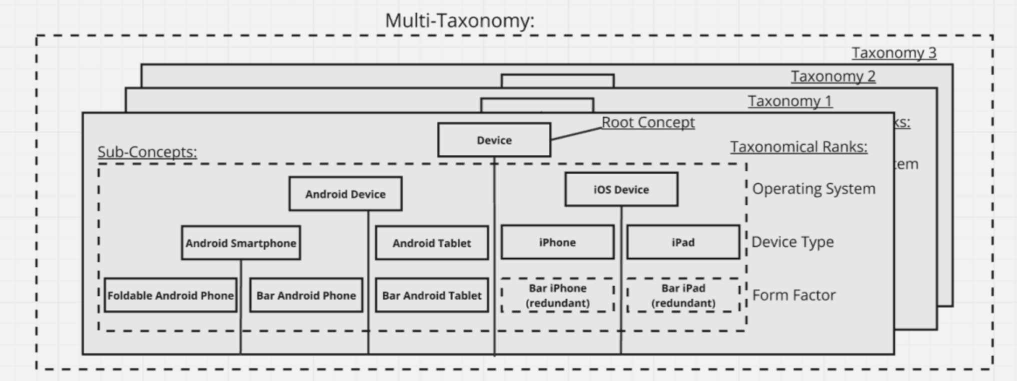

Figure 1: Multi-Taxonomy Example for the Root Concept “Device”.

4. Methodology

The methodology outlined in this section forms the core of the TaxoRankConstruct approach. This section details the steps required to implement this method, emphasizing the integration of LLM-driven processes. Given the challenges of constructing taxonomies from scratch, especially in domains where predefined hierarchies may not exist, the proposed methodology leverages the strengths of LLMs to address these challenges. The following subsections will guide you through each phase of the process, providing a detailed explanation of the techniques and strategies employed.

4.1. Initial Task Definition

The main practical task of this research is to construct the multitaxonomy T for a given root concept R and identify the set N, which encompasses all existing subconcepts and their descendant subconcepts across all levels and trees within the multitaxonomy, with N defined as the comprehensive union of all C_i within the set C.

4.2. Identifying Key Properties

To solve the problem of finding the set N of all existing subconcepts and their descendants for a root concept R in a multitaxonomy T, the first step is to determine the number of trees T_i and the depth of each tree. This is achieved by identifying the initial key properties $F_{initial}$ associated with concept R. These properties are then used to form the set F, which consists of ordered, non-overlapping subsets of $F_{initial}$. A bijection $f\colon F \overset{\leftrightarrow}{} T$ is then established, where each subset F_i from F corresponds to a tree T_i with a depth of $|F_i|$.

4.3. Iterative Concept Discovery

For each tree T_i, a set of concepts C_i is created, starting with the root concept R, which is marked as *"unexplored"*. The process involves $|F_i|$ iterations of a procedure where, for each unexplored concept $C_{ij} \in C_i$ $(0 < j \leq |C_i|)$, the set of its subconcepts M_{ij} is identified. These subconcepts are added to C_i as *"unexplored"*, and C_{ij} is marked as *"explored"*. Through $\sum_{k=1}^{|F|} |F_k|$ iterations, all subconcepts N are identified.

4.4. Finalizing Tasks

The following main tasks have been identified:

1. Determine the properties $F_{initial}$ for the root concept R in the multitaxonomy T.
2. Identify the set of property sets F for multitaxonomy T based on the properties $F_{initial}$.
3. "Determine the subconcepts M_{ij} for the concept C_{ij} in the taxonomy T_i."

   Assuming tasks (1) and (2) are resolved, a refined task is formulated:

4. Determine the subconcepts M_{ij} for the concept C_{ij}, the property F_{ik} $(0 < k \leq |F_i|)$, and the set of properties F_i in the taxonomy T_i.

4.5. LLM-Driven Taxonomy Construction

LLMs were utilized in this study to solve tasks related to taxonomy construction. These models have access to vast amounts of data and demonstrate impressive results in natural language understanding and generation, enabling them to tackle complex tasks even in a zero-shot setting. However, the quality of LLM-generated text largely depends on the context, which can significantly influence the final result. Additionally, there is an inherent element of randomness, which can cause different outputs across multiple runs.

Two primary approaches were used to interpret concepts: "as a linguist" and "as an expert." These approaches are based on two key sources of knowledge—dictionary and encyclopedic formats. Dictionary definitions provide a clear and formal structure of concepts, while encyclopedic descriptions offer broader context and cultural information. Both approaches are crucial for forming a comprehensive understanding of the properties of concept R and its related types [20].

- Example of definition generated for the Root Concept "Music": "a cultural construct varying widely among different societies based on tonal systems, scales, and patterns catering to emotional engagement;"
- Example of description generated for the Root Concept "Music": "A social phenomenon reflecting diverse traditional practices wherein communities communicate values and narratives through coordinated sonic patterns often involving singing or playing musical instruments collectively;"

4.6. Multistep LLM Processing

For tasks (1) and (2), multiple generations of descriptions and definitions of concept R were carried out using LLMs. Initially, two types of prompts were created (see Fig. 2): one to obtain definitions from the perspective of a linguist (*"Role: You are an outstanding linguist."*) and the other to obtain descriptions from an ontology expert's perspective (*"Role: You are an outstanding ontologist expert."*). Multiple generations allow for the collection of a wide range of potential definitions and descriptions, significantly improving the quality of the final result [21].

After generating descriptions and definitions, the LLM was used to extract all possible properties of R based on each text received. This resulted in a multitude of taxonomic criteria, which were then filtered. This process allows the model to filter out irrelevant properties based on the overall mass of relevant information, significantly increasing accuracy and reducing noise in the final list of properties [21].

4.7. Optimization and Finalization

Once the taxonomic criteria were extracted, the next step involved creating *"initial lists of taxonomic ranks."* This process was performed in several stages:

In the first stage, the LLM generated an ordered set of key properties (ranks) from each set of taxonomic criteria. These ranks represent the main characteristics that differentiate species within the taxonomy.

After creating k lists of ranks (corresponding to the number of sets of taxonomic criteria), the optimization stage begins. At this stage, the model is tasked with "optimizing the sets," which includes changing the order of ranks, moving them between sets, removing, modifying, and adding new ranks. This process accounts for the relationships between different properties and improves the structure and completeness of the final lists.

The output is a set of "taxonomic rank lists," which constitutes the final set F. These lists serve as the foundation for further taxonomic work, providing a more accurate and consistent representation of the relationships between concepts in the multitaxonomy.

4.8. Validation and Iteration

Task (4) for taxonomy T_i is addressed by including information about the root concept R, the set of taxonomic ranks F_i of concept R, and the taxonomic rank F_{ik} of concept C_{ij} in the prompt for generating subconcepts. Additionally, LLM is used to generate definitions for C_{ij} using R, F_i, and F_{ik} as context. For example:

Context: *"We are currently at the 'Grain pattern' level in the hierarchy (Grain pattern > Dimensional stability). The root concept of the taxonomy is Lumber wood."*

Instruction: *"Give a 50-word definition for the Grain pattern of the ontological concept 'Cross-grain' for our taxonomy."*

Model's Response: *"Cross-grain refers to a grain pattern where the wood fibers run at an angle or perpendicular to the main length, resulting in challenges for working with and reducing dimensional stability. It often leads to uneven surfaces and difficulty in machining or finishing."*

These definitions are also used as additional context when generating subconcepts.

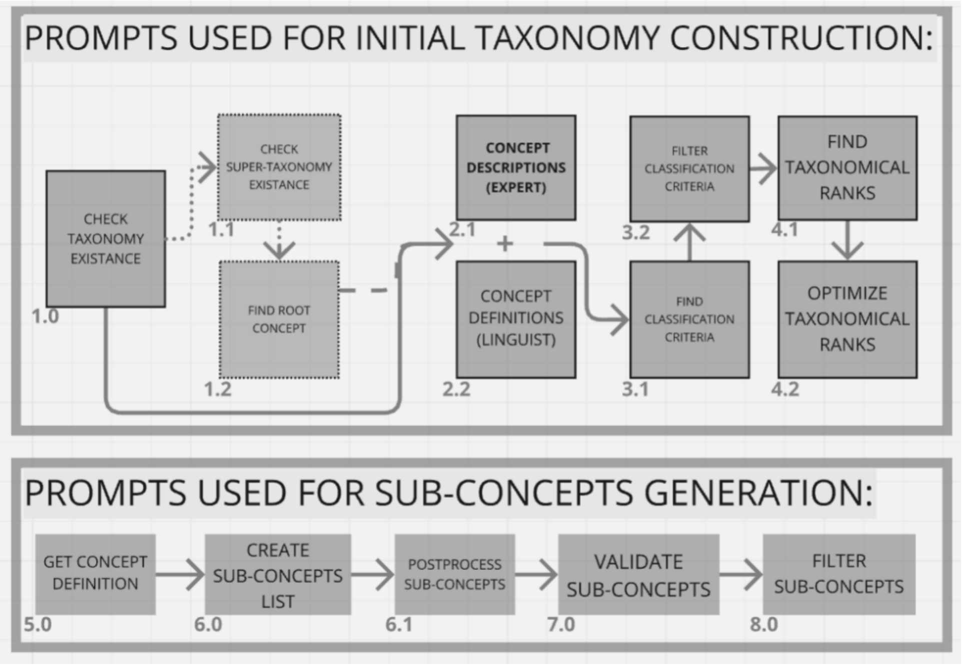
PROMPTS USED FOR INITIAL TAXONOMY CONSTRUCTION:
CHECK TAXONOMY EXISTANCE
1.0
CHECK SUPER-TAXONOMY EXISTANCE
1.1
FIND ROOT CONCEPT
1.2
CONCEPT DESCRIPTIONS (EXPERT)
2.1
+
CONCEPT DEFINITIONS (LINGUIST)
2.2
FILTER CLASSIFICATION CRITERIA
3.2
FIND CLASSIFICATION CRITERIA
3.1
FIND TAXONOMICAL RANKS
4.1
OPTIMIZE TAXONOMICAL RANKS
4.2
PROMPTS USED FOR SUB-CONCEPTS GENERATION:
GET CONCEPT DEFINITION
5.0
CREATE SUB-CONCEPTS LIST
6.0
POSTPROCESS SUB-CONCEPTS
6.1
VALIDATE SUB-CONCEPTS
7.0
FILTER SUB-CONCEPTS
8.0

Figure 2: Prompts used for the Taxonomy Construction and Sub-Concepts Generation.

The generated subconcepts are then subjected to post-processing through LLM in a Few-Shot learning format. The primary goal of this step is to prevent *"Domain Shift"*[15]. For post-processing, the model is provided with examples like: *"root concept: 'Wound', taxonomical rank: 'Location', sub-concept candidates: 'Hands', 'Knees', 'Elbows'. Provide the true sub-concepts: Wounded Hand, Wounded Knee, Wounded Elbow."* This helps maintain additional taxonomic context in the names of the subconcepts and prevents the model from deviating from the topic.

If the post-processing result successfully passes the *"validation"* — confirming that the current list of candidates is indeed an acceptable set of subconcepts for C_{ij} — the process moves to the next stage. At this stage, the model selects *"redundant subconcepts"* from the candidate set. These subconcepts are excluded, and the remaining ones form the set M_{ij}.

If the *"validation"* fails or *"redundant subconcepts"* make up more than 80% of the candidates, the attempt is considered unsuccessful, and the procedure is repeated. The maximum number of attempts is typically 5 (but can be adjusted as needed). If all attempts are exhausted, the generation is considered unsuccessful, and the concept is skipped.

This iterative process ensures that the generated subconcepts are both contextually relevant and accurately reflect the taxonomic structure, minimizing the risk of introducing irrelevant or redundant concepts into the taxonomy.

5. Experiments

In this section, we detail the experimental setup, datasets, and procedures employed to evaluate the effectiveness of the TaxoRankConstruct methodology. The primary goal of these experiments is to assess the method's ability to construct taxonomies from scratch and refine them iteratively, thereby creating coherent and meaningful hierarchical structures.

5.1. Introduction to Experimental Setup

Our experiments are designed to explore various aspects of taxonomy construction using large language models (LLMs). As previously mentioned, the concept of *multitaxonomies* is central to our approach — each taxonomy consists of multiple trees of varying depths. We structured our experiments into several scenarios, including basic multitaxonomy creation and a comparative analysis with WordNet taxonomies [22] via human evaluation. Human evaluators played a critical role in assessing the quality of the generated taxonomies, focusing particularly on the accuracy of taxonomical rank assignments and the coherence of the resulting hierarchies.

5.2. Datasets and Preprocessing

To thoroughly evaluate the TaxoRankConstruct methodology, we employed a diverse range of root concepts (*R*) from various domains. Examples of these concepts include:
'Art', 'Music', 'Transport', 'Food'
'Disease', 'Wound', 'Natural Language Processing (NLP)', 'Software'
'Artificial Intelligence', 'Organism', 'Lumber Wood', 'Electronic Component'
'Processor', 'Transistor', 'Resistor', 'Semiconductor', 'Sport'

Experiments were conducted with these root concepts and their variations, such as *'Natural Language Processing'/'NLP', 'Disease'/'Diseases', 'Organism'/'Organisms',* and *'Resistor'/'Resistors'.*

For each root concept selected in the experiments, we extracted all hyponyms from WordNet, treating them as the set of subconcepts associated with that root concept. This set of WordNet hyponyms served as a benchmark for evaluating the taxonomies generated by the TaxoRankConstruct methodology. The preprocessing steps included lemmatization and deduplication to ensure consistency and uniqueness in the evaluation set. Once preprocessing was complete, the WordNet hyponyms were combined with the subconcepts generated by TaxoRankConstruct. This combined set was then used in the human evaluation process, allowing direct comparison between our generated taxonomies and those from WordNet.

5.3. Experimental Scenarios

5.3.1. Scenario 1: Basic Taxonomy Construction

Objective:

This scenario establishes a baseline by constructing a simple taxonomy using the default settings of the TaxoRankConstruct methodology. The aim is to observe how effectively the system generates a taxonomy from a root concept and assigns taxonomical ranks to subconcepts. We generate multiple taxonomies for a single root concept and investigate the various taxonomical ranks that emerge from these taxonomies.

Procedure:

- Taxonomy Generation: The process begins by generating a diverse set of taxonomies for a given root concept using our iterative construction method. This involves verifying the root concept, generating descriptions, and assigning taxonomical criteria and ranks.
- Probabilistic Rank Generation: To address the inherent variability in model outputs, taxonomical ranks are generated multiple times. After generating multiple taxonomies for the same root concept, we compile all the taxonomical ranks that were identified across these taxonomies. The collection of taxonomical ranks can include various classification criteria, such as 'Duration', 'Type of material used', 'User interface type', and others, depending on the context of the root concept. This approach ensures a more robust set of ranks by aggregating them across iterations.
- Analysis: The next step involves an analysis of the collected taxonomical ranks. We examine the frequency and distribution of each rank, identifying which ranks are most commonly used and which are unique to specific taxonomies.
- Human Evaluation: The aggregated ranks are used to generate evaluation questions, which are then assessed by human evaluators. The question format for evaluating ranks was chosen to focus on the accuracy of highlighting important features of the root concept. The question was: *"Does the 'taxonomical_rank' accurately highlight important features of 'root_concept'?" with response options "Accurately" and "Inaccurately."*

Expected Outcome:

This baseline scenario provides a reference for evaluating the effectiveness of the TaxoRankConstruct method and sets the stage for more complex experiments.

5.3.2. Scenario 2: Comparative Evaluation with WordNet

Objective:

Compare the taxonomies generated by TaxoRankConstruct with established hierarchies from WordNet.

Procedure:

Taxonomies for various root concepts are generated and evaluated against their WordNet counterparts. In this scenario, the questions involving WordNet were framed as: *"Is '{hyponym_of_root_concept/sub-class generated}' an accepted sub-class of '{root_concept}'?" with response options "Yes" and "No."* Human evaluators assess the accuracy and relevance of these taxonomies.

Expected Outcome:
This comparison highlights the strengths and potential limitations of our approach relative to an established linguistic resource.

5.4. Evaluating Taxonomical Ranks

One of the key aspects of our methodology is the identification and evaluation of taxonomical ranks. In our experiments, we generated multiple multitaxonomies for each root concept and evaluated how the number of unique taxonomical ranks evolved across iterations. For example, we observed how the quantity and distribution of unique ranks changed with each iteration of multitaxonomy generation. From this analysis, we found that by the 10th iteration, the average percentage of unique ranks per iteration had stabilized at around 6%.

This analysis allowed us to identify the point at which additional iterations contributed minimal new information, guiding the selection of 10 iterations as the standard for further tests.

5.5. Optimizing Human Evaluation

Human evaluation was a critical component of our experimental process. Evaluators were divided into groups based on their domain expertise. For instance, concepts like '*Art', 'Music', 'Food'*, and *'Sport'* can be evaluated by individuals from general backgrounds, while more specialized concepts like *'Software', 'Electronic Component', 'Processor', 'Transistor', 'Resistor', 'Semiconductor',* and *'Lumber Wood'* need domain experts.

The root concepts selected for the primary experiments under the defined scenarios were *'Software', 'Resistor', 'Transistor',* and *'Music'*. These concepts were chosen due to their varying levels of complexity and representation in WordNet, providing a testbed for evaluating the TaxoRankConstruct methodology. By focusing on these diverse concepts, we were able to assess the methodology's effectiveness across different domains, ensuring that the results were both comprehensive and reflective of real-world applications.

Prior to formal Human Evaluation tests, we conducted numerous preliminary experiments based on subjective observations and assessments of rank quality. These experiments helped refine the methodology, tune hyperparameters, craft prompts, and select numerical parameters such as the number and maximum length of definitions, the number of subconcept generation attempts, and so on. After achieving subjectively promising results and fine-tuning the method, we finalized the parameters (which are documented in the appendix "Models") and generated the multitaxonomies for Human Evaluation tests.

Testing the quality of the generated subconcepts for a root concept like *'Software'* (which has 182 hyponyms in WordNet) requires substantial time, given that each of the 182 hyponyms would need to be evaluated against 364 questions in our chosen approach. Evaluating ranks is somewhat simpler due to the fewer questions involved.

6. Evaluation and Results

In this section, we present a comprehensive evaluation of our proposed rank-based taxonomical classification methodology using iterative construction with large language models (LLMs). The primary objective of our evaluation is to assess the accuracy, relevance, and comprehensiveness of the taxonomical classifications generated by our approach.

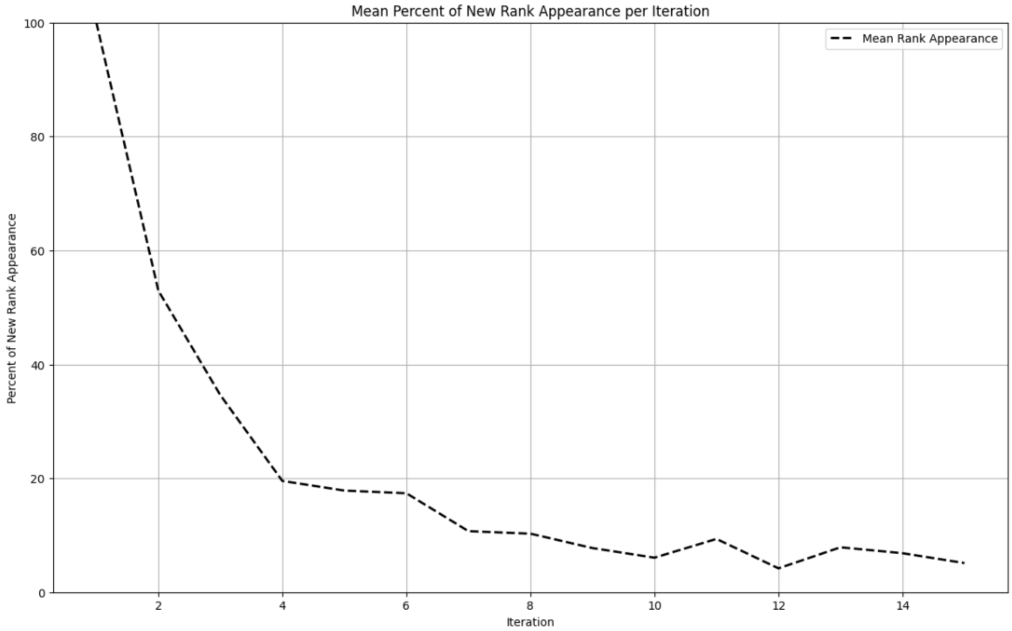

Figure 3: New Rank appearance per iteration.

6.1. Limitations of Using WordNet as a Benchmark

In evaluating the quality of the taxonomies generated by the TaxoRankConstruct methodology, we initially considered using WordNet as a benchmark due to its extensive collection of hyponyms for various concepts. However, several significant limitations prevent WordNet from serving as a reliable standard for this purpose. While numerical methods such as those discussed in [23, 24]—which involve reducing concepts to a common vocabulary—can be effective when working with predefined candidate subconcepts, they are less applicable when dealing with taxonomies generated "from scratch."

Limitations of Using WordNet:

- Inconsistent Concept Representation: WordNet presents a highly uneven distribution of hyponyms across different concepts. For example, it lists 271 hyponyms for the concept "wood," 38 for "lumber," 254 for "art," 812 for "music," 182 for "software," but only 10 for "resistor," and 5 for "artificial intelligence." This inconsistency makes it difficult to use WordNet as a reliable standard for evaluating the breadth and depth of generated taxonomies.
- Misclassification of Instances as Subconcepts: WordNet often includes instances rather than true subconcepts in its hyponym sets. For example, under "music," entries like 'colossians,' 'epistle of paul the apostle to the colossians,' and 'book of amos' appear—terms that are clearly instances or related to other domains rather than hierarchical subclasses of "music." This issue complicates the use of traditional precision, recall, f-measure, semantic overlap, and semantic cotopy metrics for evaluating taxonomy quality.
- Redundant and Non-Intuitive Hyponyms: WordNet also contains redundant hyponyms and terms that may not intuitively belong to the expected category, further distorting evaluation metrics. For instance, multiple terms that refer to the same concept (e.g., 'water-color,' 'water-colour,' 'watercolor,' 'watercolour') can artificially inflate the perceived coverage of a taxonomy. Moreover, non-intuitive hyponyms like 'apocalypse' under "music" challenge the logical coherence of the taxonomy.

6.2. The Role of WordNet in Comparative Evaluation

Despite these limitations, WordNet remains a useful reference point for evaluating the effectiveness of our taxonomy generation approach. By comparing the taxonomies generated by TaxoRankConstruct with established hierarchies derived from WordNet, we can assess the accuracy, relevance, and comprehensiveness of our taxonomies in relation to widely recognized standards. This comparative evaluation allows us to highlight the unique contributions of our methodology and identify potential areas for improvement. However, given the aforementioned issues with WordNet, this comparison is complemented by human evaluation to ensure a more nuanced and context-sensitive assessment.

6.3. Rationale for Using Human Evaluation

Given these limitations, we chose to rely on human evaluation for assessing the quality of the taxonomies generated by TaxoRankConstruct. Human evaluators are better equipped to discern the nuances of conceptual hierarchies, accurately distinguishing between true subconcepts and instances, as well as identifying and consolidating redundant terms. This approach also allows evaluators to assess whether certain hyponyms or instances, which may seem illogical out of context (e.g., 'apocalypse' under "music"), genuinely fit within the conceptual framework of the taxonomy.

Human evaluation was employed to answer key questions such as:

- Accuracy of Classification: How well do the generated taxonomical ranks represent the relationships within the taxonomy?
- Relevance and Coherence: Are the subconcepts logically organized under the root concept, and do they reflect meaningful distinctions? Are non-obvious or context-dependent terms appropriately placed?
- Identification of Redundancies and Non-Intuitive Concepts: Can human evaluators identify and reduce redundant terms in the taxonomy and flag non-intuitive or context-dependent hyponyms?

6.4. Human Evaluation

To validate our findings, we conducted a human evaluation involving domain experts and crowdworkers. We included taxonomies based on hyponym relations from WordNet in our evaluation tests, allowing us to directly compare our method against established hierarchies. The evaluation involved two types of tests: evaluating the relevance of taxonomical ranks and assessing the classification accuracy of subconcepts. Human evaluators were provided with structured questionnaires designed to test the coherence and accuracy of the generated taxonomies.

To facilitate the evaluation process, we developed an automated system for creating Google Forms via the Google Forms API, which dynamically generated evaluation forms based on the taxonomies being tested. This automation minimized manual effort and ensured consistency across evaluation tasks.

6.5. Results for Selected Concepts

For the selected root concepts *'Software,' 'Resistor,' 'Transistor,'* and *'Music,'* the evaluation results demonstrate the effectiveness of the TaxoRankConstruct methodology across different domains. The evaluation involved calculating the Average Agreement among nine domain experts, which provides insight into the consensus reached on the quality of the generated taxonomies.

- Average Agreement: The calculated Average Agreement was 0.759 for Scenario 1 (Basic Taxonomy Construction) and 0.704 for Scenario 2 (Comparative Evaluation with WordNet). These values indicate a strong level of agreement among the experts [25].
- Unique Ranks in Scenario 1: The analysis of unique taxonomical ranks for Scenario 1 revealed that the average percentage of unique ranks, which were selected by the majority of experts as *"Accurately"* representing important features of the root concepts, was 87% after the 10th iteration. The mean amount of taxonomical ranks generated per iteration is 8.9. The mean amount of ranks chosen as *"Accurately"* by the most experts is 7.3 per iteration, and the mean amount of ranks chosen as *"Inaccurately"* by the most experts is 1.6 per iteration.

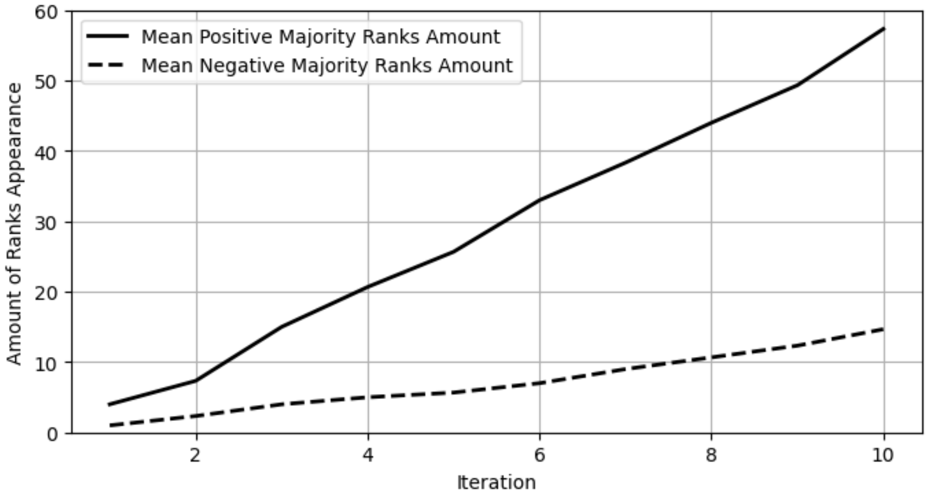

Figure 4: the Mean Amount of Ranks chosen as *"Accurately/Inaccurately."* by the Most Experts.

- Accepted Sub-classes in Scenario 2: In Scenario 2, the comparison of generated subconcepts with those from WordNet showed that the average percentage of accepted sub-classes was 79.2% for the subconcepts generated by TaxoRankConstruct, compared to 68.9% for the hyponyms derived from WordNet. This result highlights the potential of our methodology to produce accurate and contextually relevant taxonomies.

Overall, these results suggest that the TaxoRankConstruct method performs well across different domains and scenarios, offering a robust approach to taxonomy construction that is both accurate and adaptable. The higher agreement rates and improved unique rank percentages over iterations indicate that the methodology can refine taxonomies effectively, making it a promising tool for generating hierarchical structures in a variety of fields.

7. Potential Applications and Future Work

The TaxoRankConstruct method offers a novel approach to taxonomy construction using large language models (LLMs). While there are existing methods for creating taxonomies, TaxoRankConstruct allows for an iterative, rank-based process where users can select specific criteria and gradually populate the taxonomy. This approach is particularly useful for building initial taxonomic structures that can be further refined and expanded.

In this study, the primary focus has been on achieving "precision" rather than "completeness" in the results. The system performs each iteration only once and does not revisit previously processed properties, which sometimes leads to the omission of potential subconcepts. The emphasis was placed on minimizing hallucinations and irrelevant outcomes, both in terms of subconcepts and the properties themselves. Additionally, the current version of the system does not account for the fact that a taxonomy is inherently a graph rather than a simple tree or a set of trees. These limitations, including issues related to completeness, restructuring, and optimization of the placement of identified subconcepts, are planned to be addressed in future research.

At its current stage, the method supports depth-first expansion of taxonomies. Taxonomies can be exported into formats like OWL (Web Ontology Language). This basic export functionality enables users to edit the taxonomy in other tools or apply it in various applications, such as quality assessment of different NLP methods.

Looking ahead, we plan to enhance the TaxoRankConstruct tool with advanced features. These include a sophisticated export process that considers taxonomic ranks and the ability to expand taxonomies breadth-wise. These improvements will give users greater flexibility. They will also enable the creation of more comprehensive taxonomic structures. The experiments have provided valuable insights. These will guide the ongoing refinement of the methodology. We will address current limitations like taxonomy completeness and restructuring. These developments will ensure that TaxoRankConstruct remains versatile and adaptable. It will be capable of meeting the evolving needs of taxonomy construction across various domains.

8. Conclusion

In this study, we introduced a novel approach to taxonomy construction, leveraging large language models to create rank-based taxonomical classifications. Our methodology addresses the limitations of traditional taxonomy construction methods, providing a flexible and iterative framework that can adapt to various domains.

Key Contributions:

- Taxonomical Ranks, Rank-Based Classification: We developed a rank-based classification system that enhances the precision and clarity of taxonomical hierarchies. This approach ensures that classifications are based on specific, identifiable characteristics, leading to more accurate and meaningful taxonomies.
- Multi-Taxonomies: We proposed the concept of multitaxonomies, which allows for the representation of concepts through multiple hierarchical trees. This approach accommodates different perspectives and categorizations within the same domain, offering a more nuanced and comprehensive representation of concepts.
- Linguist/Expert Definitions: By incorporating definitions generated from both linguistic and expert perspectives, our method provides a rich, context-aware understanding of concepts. This dual approach ensures that taxonomical classifications are grounded in both formal and contextual knowledge.
- Few-Shot Post-Processing to Prevent Domain Shift: To enhance the relevance and coherence of generated subconcepts, we implemented a few-shot post-processing step. This technique mitigates the risk of domain shift, ensuring that the taxonomy remains consistent and contextually appropriate.

Our results demonstrate the effectiveness of the TaxoRankConstruct methodology across diverse domains. The iterative nature of our approach allows for the continuous refinement and enhancement of taxonomies, making it a valuable tool for a wide range of applications.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] C. Wang, X. He, and A. Zhou, "A short survey on taxonomy learning from text corpora: Issues, resources and recent advances," in *Proc. 2017 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2017, pp. 1190–1203. DOI: https://doi.org/10.18653/v1/D17-1123.

[2] K. R. Rao, "Taxonomy construction techniques: Issues and challenges," in *Proc. Int. Conf. Knowledge Organization*, 2011. URL: https://www.semanticscholar.org/paper/TAXONOMY-CONSTRUCTION-TECHNIQUES-%E2%80%93-ISSUES-AND-Rao/ff89917911da2a1f7a314bb1b2f033e8ec2bad0b.

[3] C. Zhang et al., "TaxoGen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering," *arXiv preprint*, arXiv:1812.09551, 2018. URL: https://arxiv.org/abs/1812.09551.

[4] J. Huang et al., "CoRel: Seed-guided topical taxonomy construction by concept learning and relation transferring," *arXiv preprint*, arXiv:2010.06714, 2020. URL: https://arxiv.org/abs/2010.06714.

[5] D. Lee et al., "TaxoCom: Topic taxonomy completion with hierarchical discovery of novel topic clusters," *arXiv preprint*, arXiv:2201.06771, 2022. URL: https://arxiv.org/abs/2201.06771.

[6] P. Mateiu and A. Groza, "Ontology engineering with large language models," *arXiv preprint*, arXiv:2307.16699, 2023. Available: https://arxiv.org/abs/2307.16699.

[7] B. Chen, F. Yi, and D. Varró, "Prompting or fine-tuning? A comparative study of large language models for taxonomy construction," *arXiv preprint*, arXiv:2309.01715, 2023. URL: https://arxiv.org/abs/2309.01715.

[8] Q. Zeng et al., "Chain-of-Layer: Iteratively prompting large language models for taxonomy induction from limited examples," *arXiv preprint*, arXiv:2402.07386, 2024. URL: https://arxiv.org/abs/2402.07386.

[9] R. Fu et al., "Learning semantic hierarchies via word embeddings," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics (ACL 2014)*, 2014, pp. 1199–1209. URL: https://aclanthology.org/P14-1113.

[10] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," *arXiv preprint*, arXiv:1511.06361, 2015. URL: https://arxiv.org/abs/1511.06361.

[11] S. Mishra, U. Sudev, and T. Chakraborty, "FLAME: Self-supervised low-resource taxonomy expansion using large language models," *arXiv preprint*, arXiv:2402.13623, 2024. URL: https://arxiv.org/abs/2402.13623.

[12] J. Shen et al., "TaxoClass: Hierarchical multi-label text classification using only class names," in *Proc. 2021 Conf. North American Chapter Assoc. Comput. Linguistics: Human Lang. Technol. (NAACL 2021)*, 2021, pp. 4251–4262. URL: https://aclanthology.org/2021.naacl-main.335.

[13] J. Huang et al., "WERECE: An unsupervised method for educational concept extraction based on word embedding refinement," *Appl. Sci.*, vol. 13, no. 22, p. 12307, 2023. DOI: https://doi.org/10.3390/app132212307.

[14] D. Budagam, S. KJ, A. Kumar, and V. Jain, "Hierarchical prompting taxonomy: A universal evaluation framework for large language models," *arXiv preprint*, arXiv:2406.12644, 2024. URL: https://arxiv.org/abs/2406.12644.

[15] M. Funk, S. Hosemann, J. C. Jung, and C. Lutz, "Towards ontology construction with language models," *arXiv preprint*, arXiv:2309.09898, 2023. URL: https://arxiv.org/abs/2309.09898.

[16] M. Lee, Z. M. Kim, V. Khetan, and D. Kang, "Human-AI collaborative taxonomy construction: A case study in profession-specific writing assistants," *arXiv preprint*, arXiv:2406.18675, 2024. URL: https://arxiv.org/abs/2406.18675.

[17] R. Amini, S. S. Norouzi, P. Hitzler, and R. Amini, "Towards complex ontology alignment using large language models," *arXiv preprint*, arXiv:2404.10329, 2024. URL: https://arxiv.org/abs/2404.10329.

[18] F. Ronzano and J. Nanavati, "Towards ontology-enhanced representation learning for large language models," *arXiv preprint*, arXiv:2405.20527, 2024. URL: https://arxiv.org/abs/2405.20527.

[19] H. Babaei Giglou, J. D'Souza, and S. Auer, "LLMs4OL: Large language models for ontology learning," *arXiv preprint*, arXiv:2307.16648, 2023. URL: https://arxiv.org/abs/2307.16648.

[20] I. Kecskes, "The interplay of linguistic, conceptual, and encyclopedic knowledge in meaning construction and comprehension," in *The Cambridge Handbook of Language in Context*, J. Romero-Trillo, Ed., Cambridge: Cambridge University Press, 2023, pp. 268–288.

[21] Y. Zhu et al., "Can large language models understand context?" *arXiv preprint*, arXiv:2402.00858, 2024. URL: https://arxiv.org/abs/2402.00858.

[22] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995. DOI: https://doi.org/10.1145/219717.219748.

[23] A. Maedche and S. Staab, "Measuring similarity between ontologies," in *Proc. EKAW'02*, Springer, 2002.

[24] P. Cimiano, A. Hotho, and S. Staab, "Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text," Jan. 2004.

[25] J. Amidei, P. Piwek, and A. Willis, "Rethinking the agreement in human evaluation tasks," preprint, 2023.

[26] D. Dvoichenkov, "TaxoRankConstruct." URL: https://github.com/supersokol/TaxoRankConstruct.

A. Online Resources

To facilitate replication and further exploration of our research, all code, prompts, parameters, and examples used in the TaxoRankConstruct methodology are available in a dedicated GitHub repository at https://github.com/supersokol/TaxoRankConstruct/

B. Models

In the TaxoRankConstruct methodology, the initialization and configuration of the large language models (LLMs) are crucial for the effective construction and iterative refinement of taxonomies. We employ three distinct models, each initialized with carefully selected hyperparameters to optimize their performance for specific tasks within the taxonomy construction process.

Verification Model - This model, based on the gpt-4o-mini architecture, is configured with a temperature of 0.90 and a top_p of 0.90, ensuring a balance between creativity and reliability. The presence penalty is set to 1.00 to encourage the generation of new content, while the frequency penalty is set to 0.00, allowing the model to freely repeat common words when necessary. This model is primarily responsible for verifying the validity and accuracy of the generated taxonomical concepts.

Re-Generation Model - Also using the gpt-4o-mini architecture, this model is configured with a higher temperature of 1.40 and a slightly lower top_p of 0.85. It features a lower presence penalty of 0.50 and a frequency penalty of 1.00, which is designed to generate diverse outputs while maintaining a moderate level of repetition control. This model is utilized for regenerating or refining concepts that need further elaboration or adjustment.

New Concept Generation Model - This model is based on the gpt-4o architecture and is configured with a temperature of 1.40, a top_p of 0.98, a presence penalty of 1.30, and a frequency penalty of 1.40. These settings are optimized to generate highly creative and varied new taxonomical concepts, which are crucial for expanding the taxonomy in novel directions.

Note that these models and their specific configurations were employed in the final stages of our experiments to optimize the balance between creativity, diversity, and accuracy in the taxonomy construction process. However, it is highly encouraged to experiment with different hyperparameters.