TaxoSurv: A Comprehensive Survey of Taxonomy Construction, Expansion, Completion, and Refinement

Zeinab Ghamlouch, Mehwish Alam

TaxoSurv: A Comprehensive Survey of Taxonomy Construction, Expansion, Completion, and Refinement

Zeinab Ghamlouch*[1], Mehwish Alam[1]

***Abstract*—Taxonomies are fundamental structures for organizing knowledge in the form of a hierarchy, supporting applications such as information retrieval, knowledge graphs, and semantic reasoning. However, many real-world taxonomies suffer from limited coverage, outdated concepts, and structural inconsistencies, motivating research on computational methods for constructing and refining hierarchical structures from heterogeneous data sources. This survey provides a systematic overview of taxonomy learning across its main tasks: construction, expansion, completion, and refinement. We introduce a structured categorization that organizes existing approaches along two dimensions, the nature of the downstream task and the methodology, including non-neural, neural, and LLM-based methods, and further analyze the benchmark datasets and evaluation protocols used across prior work. Our goal is to consolidate the state-of-the-art and propose a coherent taxonomy of methods that clarifies terminology and supports future research.**

I. Introduction

Taxonomies provide a fundamental mechanism for organizing knowledge into hierarchical structures, enabling abstraction, navigation, and semantic reasoning across domains. By structuring concepts through hierarchical "is-a" relations, taxonomies support efficient search, reasoning, and large-scale classification systems. However, real-world taxonomies are often incomplete, outdated, or structurally inconsistent. New concepts continuously emerge, domain terminology evolves, and hierarchical relations may become ambiguous or erroneous over time. Manual taxonomy engineering is costly, time-consuming, and difficult to scale [54]. These challenges have motivated extensive research into computational methods for taxonomy construction, expansion, completion, and refinement.

Over the years, a diverse range of approaches has been proposed for these tasks. Early work relied on lexico-syntactic pattern-based extraction for hypernym discovery, most notably the patterns introduced by Hearst [23], such as *"X such as Y"*, *"Y and other X"*, and *"X including Y"*, which enable the identification of hypernym–hyponym relations directly from text corpora. Subsequent research moved towards semi-supervised and web-based taxonomy induction methods [32]. With the rise of representation learning, neural and embedding-based models were introduced to capture hierarchical structure and semantic similarity more effectively [68]. More recently, Large Language Models (LLMs) have enabled prompt-based and generative strategies for taxonomy construction, expansion, and refinement [61].

Although there are existing surveys in the related areas such as knowledge graph refinement [60], LLM-enhanced knowledge representation learning [86], and domain-specific taxonomy generation [74], they typically focus on particular resources, paradigms, or application domains. To the best of our knowledge, there is currently no comprehensive survey that systematically organizes approaches across the full range of taxonomy learning tasks, i.e., construction, expansion, completion, and refinement, while providing a unified methodological classification of techniques. As a result, the field lacks a consolidated perspective that clarifies conceptual distinctions and methodological relationships across tasks.

To address this gap, this survey proposes a structured methodological categorization of computational approaches for taxonomy learning. We organise existing methods along two complementary dimensions: (i) the (downstream-)task dimension, including construction, expansion, completion, and refinement; and (ii) the methodological dimension, capturing different modeling approaches, including non-neural techniques, neural methods, and LLM-based approaches. This dual perspective enables a systematic comparison of approaches across both functional objectives and modeling principles.

Contributions

The main contributions of this survey are as follows:

- We provide a comprehensive overview of methods for taxonomy construction, expansion, completion, and refinement.
- We propose a unified methodological classification framework that systematically categorises existing approaches into non-neural, neural, and LLM-based approaches.
- We provide a systematic analysis of methodological approaches, data sources, and evaluation protocols across prior work, highlighting recurring modeling strategies and differences in task formulation and evaluation settings.
- We identify open challenges and promising research directions for scalable and adaptive taxonomy learning.

The remainder of this paper is structured as follows. Section II reviews related surveys and positions our work within the existing literature. Section III introduces the terminology and problem formulation for taxonomy learning tasks. Section IV presents the proposed methodological taxonomy of approaches. Section V discusses commonly used datasets,

∗ Corresponding Author
1 Télécom Paris, Institut Polytechnique de Paris, LTCI, Palaiseau, France.

evaluation settings and metrics. Section VI provides a comparative analysis of existing methods across the three main methodological approaches introduced earlier. Section VII outlines key challenges and future research directions. Finally, Section VIII concludes the paper.

II. Related Surveys and Positioning

Several surveys have examined related aspects of taxonomy learning, including taxonomy induction from text, knowledge representation, and hierarchical reasoning. However, existing works typically focus on specific subtasks, methodological approaches, or application domains rather than providing a unified methodological perspective across the full range of taxonomy learning tasks. For instance, Wang et al. [82] focus specifically on taxonomy construction from text corpora, organizing subtasks such as hypernym extraction and taxonomy induction.

A significant body of surveys target knowledge graph (KG) refinement and completion techniques. For example, in [60], the author provides a comprehensive overview of approaches for KG refinement, including error detection and completion strategies, along with evaluation methodologies. Complementary surveys focus on specific aspects of KG completion, such as embedding-based methods that integrate structured and unstructured information (e.g., textual, numerical, and image literals) for link prediction [18]. More recent surveys, such as Wu et al. [88] and Wang et al. [86], examine foundation models, including LLMs, and multimodal approaches, highlighting embedding-based, neural, and LLM-based techniques.

Beyond KG completion, related surveys also examine representation learning approaches for hierarchical structures. For instance, Chen et al. [13] present a comprehensive survey focusing on the importance of hyperbolic geometry for modeling hierarchical structure, which plays a key role in many neural approaches discussed in this survey (e.g., HyperExpan [37]). However, such surveys primarily address geometric representation learning rather than taxonomy learning tasks.

Other surveys concentrate on taxonomy generation within specific application domains. For instance, Spyros et al. [74] provide a comprehensive review of manual and dynamic approaches for cybersecurity taxonomy generation. Their work systematically analyzes taxonomy development techniques and datasets within the cybersecurity domain. Although such domain-specific surveys offer valuable practical insights, they are inherently limited in scope and do not provide a cross-domain methodological synthesis.

Overall, existing surveys address important adjacent problems, but they do not explicitly distinguish between taxonomy construction, expansion, completion, and refinement as separate tasks, nor do they propose a unified methodological taxonomy tailored to hierarchical "is-a" structures. This survey fills that gap by proposing a structured classification framework that disentangles downstream tasks from the proposed methodologies.

III. Terminology and Problem Formulation

A taxonomy is a hierarchical structure that organizes concepts according to semantic "is-a" (parent–child) relations. Formally, a taxonomy is represented as a Directed Acyclic Graph (DAG)

$$T = (V, E),$$

where V denotes the set of concepts (vertices) and $E \subseteq V \times V$ represents directed parent–child relations. An edge $(u, v) \in E$ indicates that concept u is a more specific concept (child) of concept v (parent).

When referring to an initial or seed taxonomy, we denote it by

$$T^0 = (V^0, E^0).$$

This structural view is standard in the taxonomy induction and expansion literature [23], [69].

In most practical settings, taxonomies are assumed to satisfy structural constraints such as acyclicity and connectivity, and may optionally enforce a single-rooted hierarchy. The primary objective of taxonomy learning methods is to infer, extend, or refine the structure T from data sources such as text corpora, structured knowledge bases, or multimodal inputs.

A. Task Formulation

Figure 1 illustrates the four tasks (taxonomy construction, expansion, completion, and refinement) through simple e-commerce examples. It shows the taxonomy before and after applying the corresponding operation, highlighting how the structure is modified under different learning settings. In the following we formalize each of the tasks.

*1) **Taxonomy Construction (Induction)**:* Taxonomy construction, commonly referred to as taxonomy induction, concerns the creation of a taxonomy when no seed taxonomy is provided [22], [32]. Given input data such as a text corpus or a knowledge base, the objective is to construct a taxonomy $T = (V, E)$ where both the concept set V and the hierarchical relations E are inferred. In some variants, a set of candidate terms may be provided, in which case the task reduces to inferring the hierarchical structure over the given concept set [67].

Formally, construction can be viewed as an optimization problem over possible edge sets,

$$\hat{E} = \arg\max_{E} \mathcal{L}(E \mid D),$$

where D denotes the input data (e.g., a text corpus, knowledge base, or term set), and $\mathcal{L}$ is a scoring function that measures how well the inferred hierarchical structure fits the data D.

In Figure 1-a, taxonomy construction is illustrated. The figure contrasts the full setting, where both the concept set and hierarchical relations are inferred from input data, with the restricted setting, where the concept set is given and only the hierarchical structure is inferred.

*2) **Taxonomy Expansion**:* Taxonomy expansion focuses on extending an existing taxonomy by integrating new concepts, which are typically added as leaf nodes, into a fixed structure. Given an initial taxonomy T^0 and a set of novel concepts C extracted from data, the expanded taxonomy is defined as

$$T = (V, E), \quad V = V^0 \cup C, \quad E = E^0 \cup R,$$

where $R \subseteq C \times V^0$ denotes the newly inferred attachment relations, i.e., each new concept is linked to a parent in the existing taxonomy, as commonly assumed in attachment-based expansion methods.

In most formulations, expansion is cast as a parent selection or attachment prediction problem: for each query concept $c \in C$, the model predicts

$$\hat{u} = \arg\max_{u \in V^0} s(u, c),$$

where $s(u, c)$ is a learned attachment score. The original structure of T^0 is preserved [49], [68], [91]. As shown in Figure 1-b, new concepts are attached to an existing taxonomy while preserving its original structure.

*3) **Taxonomy Completion**:* Taxonomy completion generalizes expansion by allowing the insertion of missing concepts at internal positions within an existing taxonomy. Given an initial taxonomy T^0 and a set of new concepts C, the goal is to insert each $c \in C$ into an appropriate position within the hierarchy.

Unlike expansion, which typically attaches new concepts as leaves, completion allows inserting concepts between existing parent–child pair, thereby modifying the internal structure of the taxonomy. The completed taxonomy is given as

$$T = (V, E), \quad V = V^0 \cup C, \quad E = E^0 \cup R,$$

where R may include relations connecting new concepts to existing nodes as well as relations that modify existing parent–child links.

Formally, for each new concept $c \in C$, the model selects a valid candidate position represented as a pair of existing nodes (u, v) such that $(u, v) \in E^0$ and u is a parent of v. The insertion operation replaces (u, v) with (u, c) and (c, v) when appropriate. In some variants, insertion may occur at the root or leaf level. Figure 1-c illustrates how new concepts can be inserted between existing parent–child pairs, modifying the internal structure of the taxonomy.

Completion is therefore commonly framed as a position ranking problem over candidate parent–child pairs [27], [95], [104]. In this sense, taxonomy expansion can be viewed as a restricted case of completion in which only leaf-level attachment is permitted.

*4) **Taxonomy Refinement**:* Taxonomy refinement aims to improve the quality, correctness, or usability of an existing taxonomy by modifying its internal structure. Unlike expansion and completion, refinement does not primarily focus on introducing new concepts, but rather on correcting erroneous relations, reorganizing abstractions, or repairing inconsistencies. Given an initial taxonomy T^0 refinement produces an updated taxonomy

$$T' = (V', E'),$$

where typically $V' = V^0$ and the edge set is modified as

$$E' = E^0 \cup E^+ \setminus E^-,$$

with E^+ denoting newly introduced relations and E^- denoting removed or corrected relations. In some settings, refinement may also modify the node set (e.g., through node merging or abstraction).

Refinement operations may include deleting erroneous edges, reassigning parent–child relations, merging overly fine-grained categories, or enforcing structural constraints such as acyclicity and consistency. As illustrated in Figure 1-d, refinement modifies the internal structure of an existing taxonomy to improve its correctness and consistency. Representative approaches include embedding-based structural correction [1], evaluation-driven refinement [38], and LLM-based validation and restructuring of hierarchical relations [61].

IV. Overall Taxonomy of Methods

The diversity of computational approaches proposed for taxonomy learning has grown substantially over time, spanning symbolic rule-based systems, statistical models, neural representation learning techniques, and more recently, LLM-based methods.

To provide a coherent synthesis of the field, we organize existing methods according to their underlying methodology rather than solely by task type. Specifically, we distinguish three major categories: (i) non-neural approaches, which rely on symbolic rules, lexico-syntactic patterns, or statistical heuristics; (ii) neural representation learning approaches, which model hierarchical relations through learned representations; and (iii) LLM-based approaches, which leverage generative and prompt-driven reasoning capabilities of large pre-trained language models.

Figure 2 presents our unified taxonomy of taxonomy learning approaches. Although the framework is defined along two complementary dimensions, downstream task and methodology, the figure organizes approaches along the methodological dimension, while the task dimension applies across all branches and is not expanded to avoid redundancy.

A comprehensive mapping of surveyed papers across taxonomy learning tasks and methodological categories is provided in Table V.

A. Non-Neural Approaches

Non-neural approaches constitute the earliest stage of taxonomy learning research, spanning from early corpus-driven pattern extraction in the 1990s to large-scale ontology engineering pipelines in the 2020s. Rather than relying on distributed representations, these methods employ symbolic patterns, statistical modeling, discrete optimization, heuristic rules, and human supervision.

1) Pattern-Based Extraction: Pattern-based extraction is one of the fundamental families of approaches for automated taxonomy induction. Hearst's seminal work [23] demonstrated that lexico-syntactic patterns (e.g., "such as", "including") can reliably extract hypernym–hyponym relations from unrestricted text corpora. This approach enabled high-precision relation discovery directly from raw text without requiring a pre-existing taxonomy.

Subsequent work by Kozareva and Hovy [32] introduced a web-scale, semi-supervised extension based on bootstrapping and graph-based positioning. They proposed Doubly-Anchored Patterns (DAPs), i.e., search query patterns of the form "<root> such as <seed> and *", where * denotes a

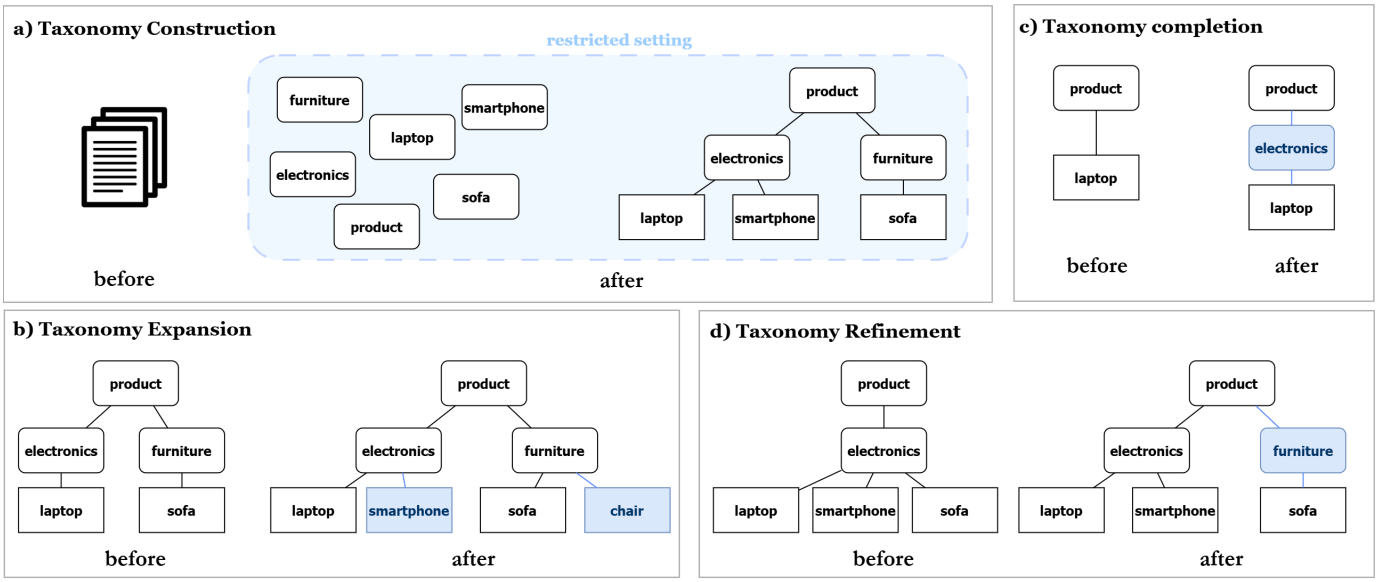

Fig. 1. Illustration of taxonomy learning tasks. Rectangular nodes denote leaf concepts, while rounded nodes denote internal concepts. Blue nodes and edges indicate newly added or modified elements between the input ("before") and output ("after") states.
For taxonomy construction, the full setting infers both the concept set and hierarchical relations from input data (e.g., text corpora). The shaded region highlights the restricted setting, where the concept set is given in advance (as candidate terms), and only the hierarchical structure is inferred.

placeholder used to retrieve new candidate terms of the same semantic type. A general root concept and a specific seed term serve as anchors, combined with concept positioning and longest-path extraction to induce domain taxonomies from scratch using web snippets. While these methods improved structural induction beyond isolated edges, they remained constrained by limited recall, as many valid relations are not expressed through explicit lexico-syntactic patterns, and by their reliance on surface expressions, making them sensitive to linguistic variation.

2) Statistical and Classical Machine Learning Methods: With the increasing availability of large-scale corpora and user interaction data, taxonomy learning began incorporating classical machine learning and probabilistic modeling techniques to move beyond purely pattern-based extraction. One of the earlier approaches by Snow et al. [73] proposes a probabilistic framework for taxonomy induction that integrates heterogeneous evidence from multiple classifiers. Specifically, the method combines a hypernym classifier based on lexico-syntactic patterns with a coordinate similarity model derived from distributional clustering. It then jointly optimizes the taxonomy structure by maximizing the likelihood of observed evidence. This formulation enables global reasoning over multiple semantic relations and leverages structural constraints to guide word sense disambiguation.

Subsequent work further explored probabilistic modeling for large-scale taxonomy construction. Wang et al. [83] introduced a hierarchical Dirichlet framework for expanding search engine taxonomies by identifying missing categories from search queries and click logs collected from user interactions. In this approach, candidate parent–child relations are scored using statistical features derived from user behavior and document distributions, and the final hierarchy is obtained via Maximum Spanning Tree (MST) optimization, which selects the globally consistent tree with the highest total edge weight.

Compared to pattern-based systems, these approaches leverage statistical evidence and learned classifiers to infer hierarchical relations, allowing them to incorporate multiple sources of evidence beyond surface text patterns. However, they remain dependent on hand-engineered features and often rely on tree-structured formulations, for instance, through MST optimization in [83], which simplifies global inference. More generally, modeling hierarchical structure remains challenging due to its inherent complexity and exponential growth, motivating the development of more expressive representations [55].

3) Graph-Based and Optimization Methods: A further methodological shift involved framing taxonomy induction as a global optimization problem over graphs. Rather than relying on greedy edge attachment, Gupta et al. [21] proposed a probabilistic framework that aggregates candidate hypernym relations into a directed graph and constructs the taxonomy by solving a minimum-cost flow optimization problem. In this formulation, edges correspond to candidate hypernym relations weighted by their confidence, and the objective is to find the lowest-cost way of assigning flow through the graph, effectively selecting paths from seed terms to root concepts. The resulting taxonomy is obtained by selecting edges that carry positive flow, yielding a globally consistent hierarchy.

More recently, Pietrasik et al. [63] proposed a non-parametric path-based model for taxonomy induction in knowledge graphs that relies on frequency-based signals, specifically the frequency of class occurrences and their co-occurrence across entities. The model formulates taxonomy induction as a discrete optimization problem over possible tree structures, where subject entities are assigned paths in the

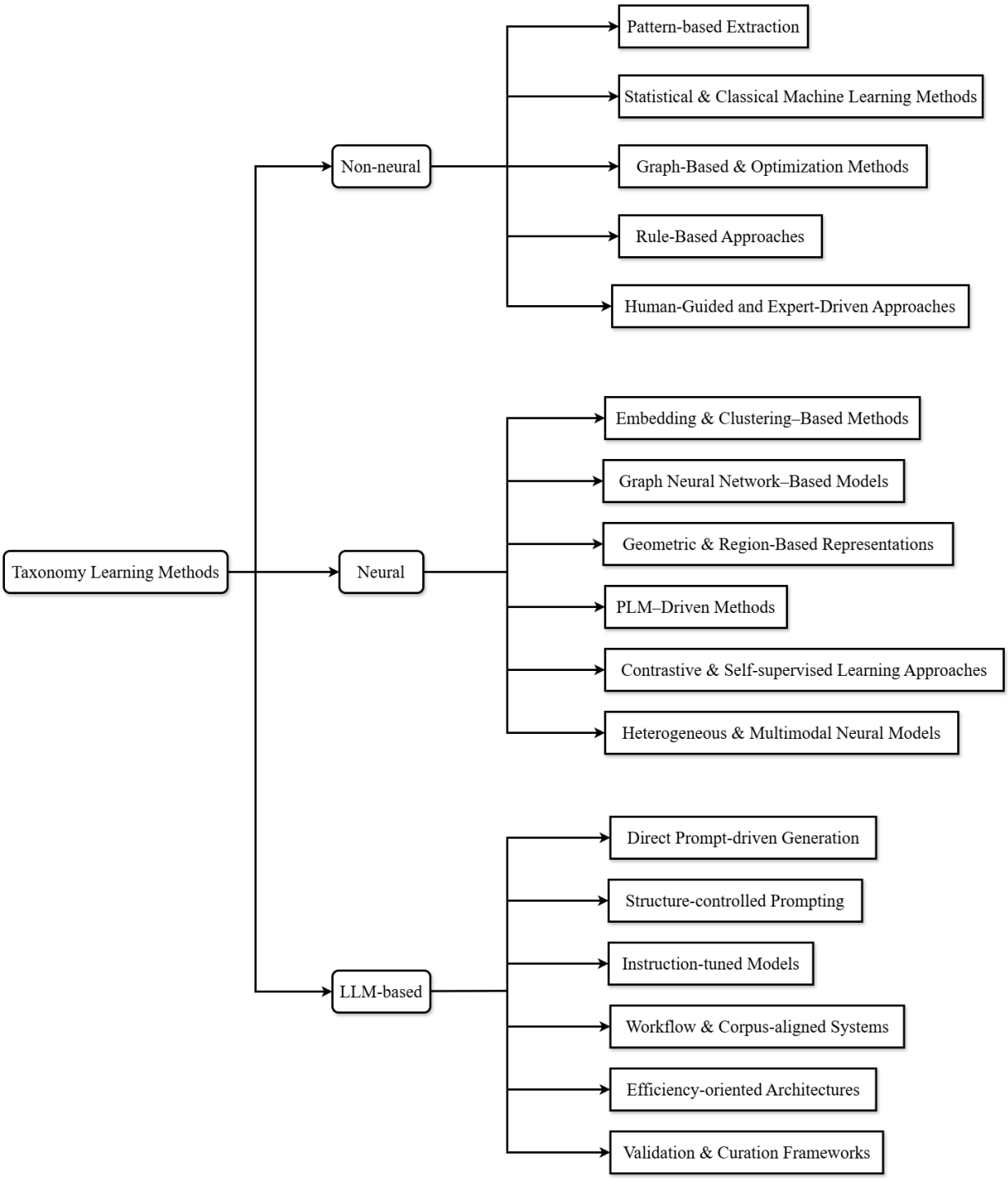

Fig. 2. Methodological taxonomy of taxonomy learning approaches, organized by methodological approaches and refined into conceptual families reflecting the evolution of the field.

hierarchy and classes are attached to nodes along these paths. The optimization is performed using simulated annealing, a probabilistic search procedure that explores the space of candidate hierarchies by iteratively proposing local modifications and accepting them based on an objective function, allowing occasional acceptance of lower-scoring solutions to escape local optima.

Compared to local edge-based methods, these approaches improve global structural consistency by jointly optimizing multiple candidate relations or structures. However, they incur higher computational cost due to global optimization and depend on the availability and quality of candidate relations or statistical signals.

4) Rule-Based Approaches: Parallel to fully automatic systems, several non-neural approaches incorporate explicit rules to guide taxonomy construction, enrichment, and refinement. These methods emphasize interpretability and controllability by explicitly modeling structural constraints.

In refinement settings, several works focus on modifying an existing taxonomy structure to improve downstream performance. Nitta [57] proposed heuristic structural operations, i.e., *promote*, *demote*, and *merge*, to iteratively restructure the hierarchy. These operations adjust the level of concepts or combine similar nodes to improve hierarchical classification performance while maintaining computational efficiency. Naik and Rangwala [53] introduced a filter-based taxonomy modification framework guided by classification performance. Although the framework itself is filter-based, the structural changes are implemented through rewiring operations. These include node creation, parent–child rewiring, and node deletion, which restructure the hierarchy based on classifier feedback and performance signals.

More recently, rule-based expert systems have been proposed to support flexible and context-dependent taxonomy adaptation. TaxoMulti [2] formulates taxonomy customization as a rule-based reasoning process for multi-channel environments where product information is distributed across diverse online and offline platforms (e.g., mobile applications, web portals, and printed catalogs). This approach addresses the *oversize* (e.g., on mobile devices) and *undersize* (e.g., on desktop interfaces) problems that arise when rigid, historically grown hierarchies are applied across different presentation contexts. The system introduces *mediator concepts*, including taxonomic dependencies that mediate between a super-concept and its sub-concepts, and taxonomic collections that mediate between a root concept and a set of super-concepts. These enable controlled restructuring of the taxonomy through explicit modification rules, allowing concepts to be combined, split, or reorganized across different levels while preserving domain semantics. The approach is implemented using a logic programming framework with a knowledge base and inference engine, enabling flexible adaptation without modifying the original taxonomy stored in the backend.

5) Human-Guided and Expert-Driven Approaches: In parallel, several approaches incorporate structured human input to guide taxonomy construction and refinement. These methods leverage human expertise either interactively or through curated knowledge resources.

Yang [96] proposed a supervised semantic distance learning framework with path-consistency control for constructing task-specific browsing taxonomies. Path consistency enforces that concepts along a root-to-leaf path remain semantically coherent and belong to the same perspective, preventing inconsistent concept sequences by minimizing semantic distances along valid paths. Furthermore, the approach incorporates interactive user guidance, where users iteratively modify the taxonomy (e.g., adding, deleting, or relocating nodes), and these edits are used as supervision to update the learned distance function and refine the hierarchy.

Crowdsourcing-based enrichment further formalizes human supervision at scale. Wang et al. [87] introduced an expertise-aware crowdsourcing framework that uses a Graph Gaussian Process to model worker expertise. By leveraging the principle of accuracy locality, i.e., that a worker's reliability varies across different regions of a taxonomy, the framework enables accurate estimation of expertise over large-scale concept spaces. Additionally, the system employs subtree recommendation to guide workers toward relevant regions of the hierarchy, reducing redundant search effort and annotation time while improving task efficiency.

Finally, large-scale ontology engineering approaches rely on expert-driven design and validation. Suchanek et al. [75] extended YAGO with a richer taxonomy by integrating a Schema.org-based upper ontology with lower-level structures derived from Wikidata via manual mappings. The resulting taxonomy is further refined through constraint enforcement, including the removal of cycles and redundant relations, as well as the application of logical constraints (e.g., domain and range restrictions and disjointness) using SHACL and OWL. This ensures structural consistency and supports reliable reasoning over the resulting knowledge base, while the lower-level taxonomy provides fine-grained, human-intelligible class distinctions.

In summary, non-neural methods span a broad design spectrum from pattern-based extraction and statistical modeling to global optimization, rule-based systems, and human-guided approaches. While they provide strong interpretability and structural guarantees, they often face limitations in recall, adaptability, and robustness to lexical variability, motivating the transition toward neural representation learning approaches.

B. Neural Representation Learning Approaches

Neural methods fundamentally reshaped taxonomy learning by replacing handcrafted linguistic cues, such as lexico-syntactic pattern-based or rule-based approaches (e.g., Hearst Patterns [23]), with distributed representations learned from data. Early work emphasized distributional similarity and clustering, but subsequent methods increasingly incorporated *structure-aware* modeling, including graph-based propagation over taxonomy neighborhoods, where concept representations are updated by aggregating information from neighboring nodes in the taxonomy graph, hierarchy-constrained geometric representations, pretrained semantic encoders, and objective

formulations based on contrastive and self-supervised learning. This progression reflects a shift from modeling isolated term similarity to jointly capturing semantic and structural properties of taxonomic hierarchies.

1) Embedding and Clustering–Based Methods: The earliest neural approaches relied on distributional embeddings combined with clustering or similarity ranking to induce hierarchical organization. Methods such as TaxoGen [103] learned adaptive term embeddings from corpora and recursively applied spherical clustering to construct topical taxonomies, such as topic hierarchies over document collections. Subsequent topic-oriented construction systems strengthened this paradigm by coupling embedding learning with hierarchical discovery of novel clusters guided by partial hierarchies [25], [34], [67]. In lexical settings, embedding-based enrichment methods ranked candidate hypernyms using cosine similarity and heuristic depth-aware penalties [65].

These approaches marked the first transition from symbolic extraction to representation-driven learning. However, because hierarchy was largely imposed *after* learning embedding (via clustering, heuristic ranking, or post-hoc constraints), they often struggled to maintain global structural consistency—motivating explicit structure-aware neural modeling.

2) Graph Neural Network–Based Models: As taxonomies are inherently graph-structured, later work introduced Graph Neural Networks (GNNs) to capture relational dependencies between nodes and to propagate hierarchical information through neighborhoods and paths. Before the adoption of GNN-based models, HiExpan [69] advanced structure-aware expansion by combining local edge predictions with a global optimization procedure that enforces consistency of the induced taxonomy structure across levels. Building on this shift toward structure-aware modeling, TaxoExpan [68] demonstrated how pseudo-supervision derived from existing taxonomies can train position-aware neural models without manual labels.

Subsequent methods deepened structural reasoning through mini-path sampling, which captures multi-level structural context by sampling short paths from the taxonomy, and multi-view learning, which integrates complementary semantic, contextual, and lexico-syntactic information [97]. They further incorporate path-based ranking with dynamic margins, which ranks candidate positions using full hierarchical paths and adapts the ranking objective based on the similarity between candidate paths [36]. In addition, explicit structural coherence is modeled via local ego-tree constraints, which enforce consistency by considering the parent node together with its ancestors and neighboring children when determining the correct placement of a concept [84]. Other work modeled implicit edge semantics, such as learning representations for parent–child pairs rather than individual nodes. This allows the models to capture asymmetric and directional properties of hypernymy relations [39], as well as transferable relational patterns, where learned structural regularities (e.g., how children attach to parents across different branches) can generalize to unseen parts of the taxonomy or new domains [66]. Taxonomy *evolution* settings further explored semi-supervised edge classification using Graph Convolutional Network (GCN)-based models to predict the validity of parent–child relations in evolving hierarchies [12].

This wave represents a major shift: the taxonomy becomes a *source of supervision* rather than just an output structure. However, it often assumes the backbone taxonomy is largely correct, so errors can propagate, and non-leaf structural edits remain difficult.

3) Geometric and Region-Based Representations: To account for the asymmetric, transitive, and exponentially expanding nature of hierarchical relations, recent work adopts non-Euclidean and region-based representations. Hyperbolic embeddings capture tree-like geometry more naturally than Euclidean space due to their exponential volume growth. For example, HyperExpan [37] leveraged hyperbolic embeddings to improve depth-aware taxonomy expansion and attachment decisions. In parallel, Aly et al. [1] showed how hierarchical structures can be embedded in hyperbolic space by representing concepts as points in a geometry where distances reflect hierarchical relations, with more general concepts placed closer to the center and more specific ones farther away, thereby preserving parent–child proximity and overall tree structure.

Region-based models represent concepts as geometric containers, where containment encodes subsumption. Box embeddings are a prominent instance of this idea [28], [105], and recent completion frameworks explicitly unify *attachment* (leaf insertion) and *insertion* (non-leaf completion) within a single region-constraint objective [95]. Other formulations introduce uncertainty and alternative partial-order encodings. For example, probabilistic and fuzzy approaches model subsumption as soft inclusion relations between concept representations, such as fuzzy set–based inclusion measures [90], while other methods model directionality and feature inheritance through non-Gaussian constraints [102]. In parallel, other frameworks introduce distributional and quantum-inspired representations, where concepts are modeled as probability distributions or states in a Hilbert space to capture uncertainty, polysemy, and multi-parent relationships in taxonomies [47], [48].

This family reflects a conceptual refinement: instead of learning representations that merely correlate with hierarchy, these methods encode partial-order properties directly into geometry. This often improves non-leaf behavior, but increases optimization complexity and sensitivity to modeling choices.

4) Pretrained Language Model–Driven Methods: The advent of Pretrained Language Models (PLMs) significantly enhanced semantic modeling capabilities. Rather than learning embeddings from scratch, these models leverage contextual encoders (e.g., BERT) to score candidate relations and positions using natural language descriptions and taxonomy-derived contexts.

PLM-based construction methods, such as [10], predict parenthood relations from flat term sets and reconcile them under global structural constraints. In expansion and completion, PLMs enable richer scoring of candidate attachments and insertion positions by encoding taxonomy structure as textual inputs. This is typically achieved by linearizing paths (e.g., root-to-node sequences), concatenating node descriptions, and incorporating local neighborhood information as context for the encoder [36]. This direction became especially

influential in *taxonomy completion*, where models rank candidate insertion *positions* rather than only predicting a single parent. Representative neural completion frameworks include triplet/pair position matching [104], structure–semantic self-supervision [27], and position-enhanced semantic matching [5]. Other approaches consider full-relation evaluation that accounts for multiple candidate parent–child relations rather than only leaf attachments, mitigating the bias toward placing new nodes at the leaves of the taxonomy [85]. Additional methods incorporate relation-aware mutual attentions over term and definition contexts [101], as well as multi-task prompt-learning as a supervised-style cross-encoder, which jointly encodes the query and candidate parent–child pairs using prompt-based formulations and learns multiple related decisions (e.g., parent, child, and attachment) within a unified framework [92]. Beyond relation prediction, completion has also been extended to *implicit concept insertion*, where concepts are inserted without explicitly predicting a single parent–child link, often involving the identification or generation of intermediate internal nodes [71], [100].

PLM-driven methods improve semantic generalization and reduce brittle feature engineering, especially in low-resource and cross-domain settings. However, they often rely on high-quality descriptions and still require structural reconciliation to ensure valid hierarchies.

5) Contrastive and Self-supervised Learning Approaches: More recent research explores the design of training objectives alongside model architectures. Contrastive and self-supervised learning reshape representation spaces so that semantic similarity aligns better with hierarchical constraints. Representative contrastive structure learning frameworks directly optimize global structural coherence by constructing positive pairs from valid hierarchical relations (e.g., parent–child or ancestor–descendant pairs) and contrasting them against hard negative pairs, encouraging representations that respect the global organization of the taxonomy [43], while completion-oriented approaches exploit multiple structural "views" such as ancestor, descendant, and sibling contexts to define positives and hard negatives. In this setting, the correct parent–child position is treated as a positive example, whereas structurally plausible but incorrect positions, for example under sibling or nearby nodes, are treated as hard negatives [58].

Similarly, many modern completion systems generate supervision directly from the taxonomy itself through pseudo-sentences, masked relation recovery, negative sampling, or retrieval-based hard negative mining [27], [85], [100]. In settings with limited labeled data, parameter-efficient adaptation further operationalizes self-supervised learning by leveraging supervision derived from the taxonomy structure and transferring shareable hierarchical knowledge from high-resource taxonomies to fine-tune pretrained models on small labeled edge sets [93]. Overall, the key innovation in this family is redefining supervision: hierarchical structure provides scalable training supervision that reduces reliance on costly annotation.

6) Heterogeneous and Multimodal Neural Models: Finally, neural taxonomy learning has expanded beyond purely textual semantics. In applied domains such as e-commerce and labor market analytics, user interaction traces and domain corpora provide implicit supervision for enrichment and evolution [6], [14], [20], [40]. Topic-oriented systems combine document-level modeling and hierarchical phrase/cluster discovery to align taxonomies with corpora [34], [67]. Multimodal approaches incorporate visual information, such as images associated with concepts, by learning joint text–image representations that improve concept grounding and disambiguation when textual names or descriptions are insufficient [106].

These models demonstrate the adaptability of neural taxonomy learning across heterogeneous data. However, they often assume a stable backbone taxonomy (i.e., a seed structure), while large-scale structural validation or correction is typically handled by separate refinement pipelines.

C. Large Language Model–Based Approaches

The emergence of LLMs has introduced a different approach to taxonomy learning. Rather than explicitly learning structural representations from data, LLM-based methods rely on prompt-based generation, in-context learning, and structured interaction to infer, validate, and reorganize hierarchies.

Since 2023, this paradigm has evolved rapidly. The progression can be understood as a sequence of methodological waves: (i) direct prompt-driven hierarchy generation, (ii) structure-controlled prompting with filtering mechanisms, (iii) instruction-tuned taxonomy reasoning, (iv) workflow-oriented and corpus-aligned systems, and (v) scalability and refinement-focused frameworks. We organize the literature into six families reflecting these shifts.

1) Prompt-Driven Hierarchical Generation: The earliest LLM-based taxonomy systems treated hierarchical inference as a direct generation problem. Given a root concept, a candidate term, or a partial hierarchy, the model was prompted to produce hypernyms, parent nodes, or levels of the hierarchy with minimal structural control. This paradigm is exemplified by FLAME [49], which leverages few-shot taxonomy-aware prompting in low-resource settings and relies on in-context demonstrations to guide parent prediction. In addition, FLAME incorporates reinforcement learning, specifically Proximal Policy Optimization (PPO), to fine-tune the model by aligning generated outputs with lexical and semantic reward signals derived from true hypernym relations.

Similarly, sequence-to-sequence formulations such as TaxoSeq [76] cast taxonomy construction as a structured generation problem by linearizing tree structures into bracketed sequences and predicting insertion positions through sequence generation. Constrained decoding ensures syntactically valid tree outputs, while a post-inference sequence complement module is used to correct structural errors such as missing brackets or invalid tokens.

These approaches show that LLMs encode latent hierarchical knowledge and can generate coherent taxonomies with minimal supervision. However, they exhibit key weaknesses like hallucinations, inconsistent depths, structural violations, and prompt sensitivity which worsen at scale, motivating mechanisms for structural consistency.

2) Structure-Controlled Prompting and Filtering: To mitigate hallucination and improve structural consistency, subsequent work introduced external control mechanisms, such as

ranking filters, similarity-based pruning, top-k candidate selection, and constrained decoding, layered on top of prompting. Rather than relying solely on raw LLM outputs, these systems filter and refine candidate relations before or after generation.

Chain-of-Layer [98] integrated ensemble ranking filters to suppress implausible edges. CodeTaxo [99] combined code-language prompting with semantic similarity filtering to restrict candidate parents. Other systems adopted top-k pruning and staged retrieval before LLM scoring. The central idea in this wave is that structure must be enforced *externally*, since LLMs do not inherently guarantee hierarchical validity.

While effective, these pipelines rely on prompting with external filtering rather than models trained for taxonomy-specific reasoning, motivating instruction-tuned approaches aligned with taxonomy tasks.

3) Instruction-Tuned and Task-Aligned Taxonomy Reasoning: Due to limitations of prompt-based approaches, including sensitivity to prompt design, reliance on external filtering mechanisms, and the absence of task-specific training for taxonomy reasoning, subsequent work focused on aligning LLMs with taxonomy-specific operations through instruction tuning and lightweight adaptation.

TaxoLLaMA [50] demonstrated that WordNet-guided instruction tuning enables a unified model capable of performing multiple taxonomy-related tasks under a shared hypernym reasoning interface. A unified taxonomy-guided instruction tuning framework further showed that sibling-finding and parent-finding tasks can be jointly learned under a shared supervised training objective across hierarchical tasks [70].

This line of work replaces prompt design and external filtering with instruction-tuned models for taxonomy tasks, enabling consistent handling of operations such as parent prediction and sibling identification within a single model.

4) Workflow-Oriented and Corpus-Aligned Frameworks: As practical deployment demands stability and adaptability, later systems embedded LLMs within broader workflows. These workflows integrate corpus-derived information such as document-level co-occurrence statistics, topic distributions, semantic similarity scores, clustering, and iterative refinement.

TaxoAdapt [31] aligned multidimensional research taxonomies, such as scientific topic hierarchies spanning multiple domains or facets, with evolving scientific corpora, combining LLM-based classification with clustering and corpus-based relevance measures to reduce drift over time. Knowledge materialization approaches [24] extracted large volumes of LLM-expressed knowledge into persistent knowledge bases and induced taxonomic structure over the resulting set of extracted class-level concepts (i.e., candidate classes derived from LLM-generated triples such as `instance_of` relations). Collaborative systems such as TAXMAP [15] leveraged consensus-style mapping and external similarity measures to enrich existing taxonomies with improved precision. Interactive hybrid systems [64] further integrated topic modeling and LLM reasoning with expert-in-the-loop validation.

These approaches emphasize workflow-level integration, where hierarchical quality emerges from the combination of corpus alignment, structured processing, and human oversight rather than from single-step generation.

5) Scalability and Efficiency-Oriented Architectures: As taxonomy size grows, standard LLM inference becomes computationally expensive due to context length limits and candidate enumeration. Recent work therefore focuses on scaling mechanisms that optimize both representation and retrieval.

COMI [94] addressed this challenge by compressing concept representations into compact, reusable embeddings. For example, it prompted the LLM to summarize a concept into a single-token representation and extracted its hidden vector as a task-specific embedding. These representations were stored and reused during inference, allowing the model to process longer taxonomy paths without repeatedly encoding full textual descriptions, thereby reducing computational cost.

Other approaches improved scalability through task decomposition and candidate filtering. Chain-of-Layer [98] reformulated taxonomy induction as a layer-wise process, iteratively selecting and validating candidate relations at each level using ranking-based filtering to limit error propagation. CodeTaxo [99] improved efficiency by reformulating taxonomy expansion as a code completion task and applying semantic similarity-based retrieval to restrict the set of candidate entities included in the prompt, reducing both token usage and search space.

In summary, these approaches achieve scalability by combining representation compression, staged retrieval, and candidate pruning, enabling LLM-based taxonomy systems to operate efficiently on large-scale hierarchies.

6) Validation, Repair, and Semi-Automatic Curation: A parallel line of work repositions LLMs as validators and curators rather than primary generators. In this setting, structural guarantees are enforced through graph operations, heuristics, and human supervision.

WiKC [61] refined the Wikidata taxonomy by combining per-edge LLM judgments with graph-mining operations to remove incorrect or redundant relations and rewire validated links. Semi-automatic abstraction methods [51] used LLM reasoning to merge overly granular categories while preserving downstream classification performance. For example, Taxoria [19] generated candidate classes via prompting and integrated them into existing taxonomies through validation and provenance tracking. These approaches highlight that LLM-based taxonomy learning benefits from external structural constraints and human oversight, particularly in large-scale, real-world deployments.

Across these developments, LLM-based taxonomy learning evolved from direct prompting to structure-controlled, instruction-aligned, and scalable systems. Compared to neural approaches, it emphasizes flexible reasoning and rapid adaptation, but structural validity and reproducibility often rely on external filtering or alignment, highlighting challenges in ensuring hierarchical soundness.

V. Benchmark Datasets and Evaluation Setting

Beyond methodological diversity, taxonomy learning research varies widely in data sources, benchmark datasets, evaluation tasks, and metrics. To provide an empirical grounding, we summarize quantitative trends from our curated collection

of papers for which evaluation details (datasets, tasks, and metrics) are explicitly reported[1].

A. Datasets and Evaluation Resources

In this subsection, we organize commonly used resources by dataset family and focus on how they are used in practice, highlighting representative evaluation setups, typical metrics, and recurring strengths and limitations across studies. Table I summarizes key statistics of the most commonly used benchmarks discussed below.

a) SemEval Benchmarks: The SemEval taxonomy benchmarks [8], [30] are among the most widely used evaluation resources in taxonomy learning research, particularly for controlled comparison across methods. The most commonly used datasets originate from the SemEval-2016 shared tasks (Tasks 13[2] and 14[3]), including the *environment*, *science*, and *food* domain taxonomies curated for evaluation. These datasets provide relatively compact, human-constructed hierarchies that enable standardized experimentation under reproducible settings.

In the surveyed literature, the SemEval taxonomies are used primarily for **taxonomy expansion** under a seed-taxonomy protocol, commonly referred to as the *held-out node protocol*, where part of the gold hierarchy is retained while a subset of nodes is removed and used as test instances; models are then evaluated based on their ability to correctly reattach these nodes to the remaining taxonomy. Representative neural methods include STEAM [97], TEMP [36], BoxTaxo [28], and FUSE [90], which typically formulated expansion as parent prediction using ranking-based evaluation. Prompt- and LLM-based approaches follow similar protocols: TaxoPrompt [91], TaxoSeq [76], CodeTaxo [99], TEF [44], and recent hybrid frameworks such as LORex [46] all evaluated on SemEval under held-out node insertion settings. Across these works, evaluation typically relies on accuracy, Mean Reciprocal Rank (MRR), and Wu & Palmer similarity, reflecting the ranking-based formulation of the task.

SemEval datasets are also used for **taxonomy completion**, although less frequently and under more varied setups. For example, ATTEMPT [89] evaluated completion through a two-stage process combining parent prediction and child labeling, while RAMA [101] framed completion as joint parent–child position identification. In addition, the SemEval-2016 Task 14 dataset supports a distinct completion formulation centered on WordNet enrichment via Out-Of-Vocabulary (OOV) term attachment; systems such as TALN [4] and Deftor [79] attach unseen terms to WordNet synsets using similarity between textual definitions (glosses) of terms and candidate WordNet synsets.

A smaller subset of studies use SemEval for **taxonomy construction** or reconstruction from flat vocabularies. For instance, Gupta et al. [21] and Chen et al. [10] evaluated induced taxonomies against SemEval gold hierarchies using edge-level precision, recall, and F1, highlighting its role beyond expansion.

The popularity of SemEval stems from its interpretability, moderate scale, and shared evaluation protocols, which enable direct comparison across symbolic, neural, and LLM-based approaches. However, several limitations are evident across the literature. First, the same datasets are reused under different task formulations (construction, expansion, completion), which complicates cross-paper comparison. Second, many works assume a clean seed taxonomy and focus primarily on leaf insertion (e.g., Yu et al. [97]; Liu et al. [36]), simplifying real-world scenarios where structural errors and non-leaf insertion are common. Third, performance often depends strongly on definitional or contextual text quality (e.g., Wang et al. [84]; Zeng et al. [99]), making improvements partly data-dependent. Finally, the relatively small size and domain specificity of SemEval taxonomies limit conclusions about scalability and robustness in large, noisy taxonomies.

b) WordNet-based Benchmarks.: WordNet [45] and its derived subsets constitute one of the most widely used evaluation resources across taxonomy learning tasks, particularly for expansion and completion. Unlike SemEval, which provides domain-specific hierarchies curated for shared-task evaluation, WordNet offers a large-scale, manually constructed lexical taxonomy with broad coverage and deep hierarchical structure. As a result, WordNet-based benchmarks, often constructed from WordNet sub-taxonomies [7], are frequently used to evaluate scalability, generalization, and robustness across both neural and LLM-based approaches.

In the surveyed literature, WordNet is most commonly used for **taxonomy expansion** under held-out node insertion protocols, where a subset of nodes is removed from the taxonomy and models are evaluated based on their ability to recover the correct parent for reinsertion. Many representation-learning approaches evaluate expansion by removing a subset of nodes from WordNet subtrees and measuring whether models can correctly recover the ground-truth parent of each removed node (or rank it highly among candidate parents) for reinsertion into the remaining seed taxonomy. Representative methods include BoxTaxo [28], DNG [102], QuanTaxo [47], and TaxoBell [48], which framed the task as ranking candidate parents using embedding-based scoring. Prompt- and LLM-based approaches adopt similar evaluation setups: TaxoPrompt [91], CodeTaxo [99], and LORex [46] all evaluated WordNet expansion using ranking protocols over candidate parent sets. Across these works, evaluation typically relies on ranking-based metrics such as Mean Reciprocal Rank (MRR), Hits@K, and Mean Rank (MR), reflecting the large candidate space induced by WordNet's size.

WordNet is also widely used for **taxonomy completion**, often in more structurally complex settings. For example, TMN [104] evaluated completion by ranking candidate parent–child position pairs for inserting new concepts into WordNet noun and verb hierarchies. Subsequent models such as QEN [85], TaxoComplete [5], TacoPrompt [92], and COMI [94] extended this setup with richer structural modeling or LLM-based reasoning. In these works, WordNet serves as a large-scale benchmark for evaluating non-leaf insertion and

[1]Categories are non-exclusive: a paper may use multiple datasets and report multiple metrics.

[2]https://alt.qcri.org/semeval2016/task13/

[3]https://alt.qcri.org/semeval2016/task14/

global hierarchy reasoning beyond simple parent prediction.

Beyond expansion and completion, WordNet is also used in more specialized scenarios. Earlier embedding-based work by Sand et al. [65] evaluated lemma insertion into Norwegian WordNet using exact and soft attachment accuracy. TEAM [62] evaluated multilingual WordNet expansion with attach-and-merge operations across Assamese, Bengali, and Hindi WordNets. These uses highlight WordNet's flexibility across languages and task formulations.

The widespread use of WordNet stems from several advantages. Its scale and structural depth make it suitable for testing hierarchical reasoning under realistic candidate spaces, while its manual construction ensures relatively high-quality gold relations. However, the literature also highlights several limitations. First, WordNet-based experiments often rely on artificial held-out splits that may not reflect real-world taxonomy evolution. Second, performance frequently depends on textual definitions or pretrained embeddings aligned with WordNet vocabulary. Third, WordNet focuses on lexical semantics, which may limit its representativeness for domain-specific or non-linguistic taxonomies such as skills or products. As a result, strong performance on WordNet does not always translate to downstream applications in specialized domains.

c) MAG-based Benchmarks.: Taxonomies derived from the Microsoft Academic Graph (MAG) [72], particularly MAG-CS [68] and MAG-PSY [104], are frequently used to evaluate taxonomy expansion and completion in domain-specific scientific settings. Unlike WordNet, which represents general lexical knowledge, MAG-based taxonomies capture field-of-study hierarchies constructed from scholarly metadata, making them suitable for testing methods under realistic domain distributions and large candidate spaces.

In taxonomy expansion, MAG-CS and MAG-PSY are commonly used under held-out node protocols similar to WordNet. Representative approaches include TaxoPrompt [91], DNG [102], QuanTaxo [47], and TaxoBell [48], which frame expansion as ranking candidate parents using semantic and structural signals. These works typically evaluate performance with ranking-based metrics such as MRR, Hits@K, and MR, reflecting the large number of candidate attachment points in MAG hierarchies.

MAG-based taxonomies are also widely used for taxonomy completion. For example, TMN [104] evaluated completion by ranking candidate parent–child position pairs for inserting new concepts into MAG-CS and MAG-Psychology. Subsequent models including TaxoEnrich [27], TaxoComplete [5], TacoPrompt [92], and TAXBOX [95] adopted similar evaluation setups while incorporating richer structural or semantic modeling. In these works, MAG datasets enable testing beyond lexical semantics by capturing domain-specific terminology and deeper hierarchies.

MAG-based benchmarks offer several advantages. Their relatively large scale and domain grounding make them more representative of real-world taxonomy construction scenarios than smaller curated benchmarks. However, several limitations are also noted in the literature. First, MAG taxonomies are often noisier and less consistently curated than WordNet, which can affect evaluation reliability. Second, experiments frequently depend on pretrained embeddings derived from external corpora aligned with MAG terminology. Finally, MAG benchmarks are domain-specific, which may limit cross-domain generalization of learned models.

d) Domain-specific and Custom Taxonomies.: Several works evaluate taxonomy expansion or enrichment on application-driven resources. For instance, Butt et al. [9] constructed a domain-specific automotive skills dataset from job postings to study dynamic taxonomy expansion and completion under changing real-world data distributions. In a different setting, Qu et al. [64] demonstrated an interactive human-in-the-loop expansion framework using customer-agent dialogues from the *ABCD* (Action-Based Conversation Dataset) [11] and biomedical literature mapped into MeSH hierarchies, emphasizing qualitative usability rather than standardized quantitative evaluation.

Domain-specific datasets are also widely used for taxonomy completion. ICON [71] evaluated implicit concept insertion on large e-commerce taxonomies (eBay and AliOpenKG), while COMI [94] and CoSTC [58] evaluated completion on MeSH and SemEval-derived food taxonomies under ranking-based setups. In topic-oriented settings, TopicExpan [34] and TaxoCom [33] constructed and completed topic taxonomies using document corpora such as Amazon reviews, DBpedia, arXiv, and the *New York Times*, evaluating cluster-level coherence and structural consistency rather than relation-level accuracy.

In taxonomy refinement, custom datasets are even more common. Some works focus on improving alignment between the taxonomy and real-world domain structure. For example, TaxoRef (Malandri et al. [38]) refined the ESCO occupation taxonomy using large-scale job advertisement corpora, while Peng et al. [61] refined Wikidata's class hierarchy using LLM-based validation to reduce redundancy and improve structural consistency. Other works address refinement through application-driven objectives, where the taxonomy is modified to optimize downstream performance. For instance, some studies improve hierarchical classification performance on web taxonomies (Naik and Rangwala [53] & Nitta [57]), while others customize product taxonomies for multi-channel e-commerce deployment (Angermann [2]). These studies typically evaluate refinement using downstream metrics such as classification accuracy, F1, or structural consistency rather than direct comparison against a gold hierarchy.

Domain-specific benchmarks reflect realistic data distributions and support downstream evaluation, but are often proprietary, weakly standardized, and evaluated under task-specific protocols, limiting cross-study comparison. As a result, they are less frequently used for controlled benchmarking than SemEval, WordNet, or MAG.

Taken together, these benchmarks serve complementary roles. SemEval provides controlled and interpretable evaluation, but their relatively small size and simplified protocols limit conclusions about scalability. WordNet-based benchmarks emphasize large-scale hierarchical reasoning and remain central for expansion and completion evaluation, though they primarily reflect lexical semantics rather than domain-specific structures. MAG offers a balance between scale and domain

specificity, though it often introduces noise and dependence on external embeddings. Domain-specific taxonomies better reflect real-world deployment but lack standardization. Overall, the absence of unified benchmarks complicates cross-paradigm comparison and highlights the need for more standardized and reproducible evaluation frameworks.

TABLE I
STATISTICS OF COMMONLY USED TAXONOMY BENCHMARKS. $|N|$ DENOTES THE NUMBER OF NODES, $|E|$ THE NUMBER OF EDGES, AND $|D|$ THE TAXONOMY DEPTH. STATISTICS ARE REPORTED FROM REPRESENTATIVE BENCHMARK DESCRIPTIONS IN PRIOR WORK.

Family	**Dataset**	$\|N\|$	$\|E\|$	$\|D\|$
SemEval	Environment	261	261	6
	Science	429	452	8
	Food	1486	1576	8
WordNet	WordNet-Noun	83,073	76,812	20
	WordNet-Verb	13,936	13,403	13
MAG	MAG-CS	24,754	42,329	6
	MAG-PSY	23,187	30,041	6
Domain-specific	MeSH	9,710	10,498	12

B. Evaluation Metrics

Evaluation practices in taxonomy learning vary substantially across tasks and methodological paradigms. Nevertheless, a small set of metrics consistently dominates evaluation across the literature. Based on our analysis, the most frequently reported measures include precision, recall, F1-score, accuracy, MRR, Hits@K, and MR. Broadly, these metrics reflect two dominant evaluation formulations: classification-style evaluation for relation prediction and reconstruction tasks, and ranking-based evaluation for attachment or candidate selection settings.

Tables II, III and IV summarize representative results for taxonomy expansion, completion, and construction methods across multiple benchmarks. Most results are drawn from original papers, with a few adopted from prior comparative studies. Results are not always directly comparable due to differing experimental settings.

In contrast, we do not include a dedicated table for taxonomy refinement methods. Unlike other tasks, refinement approaches are often evaluated as post-processing steps applied to different input taxonomies, and the lack of a unified benchmark or shared baseline makes direct comparison across studies difficult.

a) Precision, Recall, and F1-score.: Precision, recall, and F1-score are widely used in taxonomy learning, particularly in settings where the task is framed as hypernym prediction, edge classification, or taxonomy reconstruction.

These metrics have been widely used across both classical and neural approaches. For example, [32] evaluated reconstructed WordNet sub-taxonomies using precision and recall over induced *is-a* relations. In shared-task settings, [4] reported recall and F1 when attaching OOV terms to WordNet in SemEval-2016 Task 14. More recent neural approaches also rely on these metrics; for instance, [103] evaluated topical taxonomy construction using relation-level accuracy and complementary clustering measures, while F1-based evaluation remains common across supervised and semi-supervised formulations. In contrast, LLM-based systems such as [50] primarily reported F1 alongside Average Precision across multiple taxonomy tasks.

b) Accuracy.: Accuracy is frequently reported in formulations where taxonomy learning is treated as a classification or parent prediction problem. It measures the proportion of correctly predicted relations or attachment decisions among all predictions.

Several neural and LLM-based methods reported accuracy for controlled candidate selection tasks. For instance, [49] evaluated parent prediction accuracy under constrained candidate sets, while [9] reported accuracy alongside other classification metrics. More recently, [99] evaluated taxonomy expansion accuracy across WordNet, Graphine [35], and SemEval benchmarks under a code-completion prompting formulation.

c) Mean Reciprocal Rank (MRR).: MRR is widely used in neural and embedding-based approaches that frame taxonomy learning as a ranking problem. It evaluates the average reciprocal rank of the correct parent (or edge) among candidate predictions, rewarding systems that rank the correct answer highly.

Many representation learning approaches adopt MRR for attachment evaluation. For example, [68] reported MRR over WordNet-based benchmarks, while [105] used MRR to evaluate candidate ranking quality. Similarly, [50] reported MRR-based evaluation across multiple taxonomy tasks when candidate ranking is available, although such metrics are generally more applicable to ranking-based models than to purely generative formulations.

d) Hits@K / Recall@K.: This metric measures whether the correct relation appears among the top-K predictions and is particularly common in neural and ranking-based formulations where systems output ranked candidate lists.

For instance, [37] reported Recall@K under held-out node settings, while [41] evaluated expansion performance using Recall@K across candidate parent sets. More recent LLM-based approaches such as [94] also reported Recall@K and Hits@K for taxonomy completion framed as ranking over candidate parent–child positions. These metrics are often reported alongside MRR to provide a complementary perspective on ranking quality.

e) Mean Rank (MR).: Mean Rank evaluates the average rank position of the correct answer among candidates. Unlike MRR, which emphasizes high-ranking predictions, MR is sensitive to outliers and penalizes lower-ranked correct predictions more heavily.

Several neural and LLM-based approaches report MR alongside MRR and Hits@K for comprehensive ranking evaluation. For example, [68] and [37] reported MR over WordNet-based benchmarks, while [94] adopted MR for taxonomy completion experiments.

f) Other Metrics.: Beyond these dominant measures, several task-specific and less frequently used metrics appear across the literature. Structural similarity measures such as

TABLE II
REPORTED RESULTS OF REPRESENTATIVE TAXONOMY EXPANSION METHODS ACROSS COMMONLY USED BENCHMARKS. "–" INDICATES THAT THE METRIC WAS NOT REPORTED OR NOT APPLICABLE. RESULTS MAY COME FROM DIFFERENT PAPERS/SETTINGS AND ARE THEREFORE NOT ALWAYS DIRECTLY COMPARABLE.

	SemEval-Env			SemEval-Sci			SemEval-Food			WordNet		Graphine		MAG-CS		MAG-PSY		WordNet-Noun		WordNet-Verb	
Method	Acc	MRR	Wu&P	Acc	MRR	Wu&P	Acc	MRR	Wu&P	Acc	Wu&P	Acc	Wu&P	MRR	F1	MRR	F1	MRR	F1	MRR	F1
TaxoExpan [68]	0.111	0.323	0.548	0.278	0.448	0.576	0.276	0.405	0.542	0.198	0.648	0.245	0.659	0.692	0.196	0.775	0.294	0.562	0.199	0.439	0.124
STEAM [97]	0.361	0.469	0.696	0.365	0.483	0.682	0.342	0.434	0.670	0.232	0.624	0.203	0.631	–	–	–	–	–	–	–	–
TEMP [36]	0.492	0.635	0.777	0.578	0.675	0.853	0.476	0.605	0.810	0.294	0.657	0.359	0.738	–	–	–	–	–	–	–	–
HEF [84]	0.553	0.653	0.714	0.536	0.627	0.756	0.479	0.555	0.735	0.164	0.603	0.255	0.665	–	–	–	–	–	–	–	–
BoxTaxo [28]	0.381	0.471	0.754	0.318	0.453	0.647	–	–	–	0.264	0.639	0.292	0.682	–	–	–	–	–	–	–	–
PEB-TAXO [105]	0.481	0.529	0.771	0.424	0.504	0.731	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
FUSE [90]	0.423	0.583	0.776	0.399	0.529	0.734	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
TaxoPrompt [91]	0.574	0.684	0.836	0.614	0.687	0.856	0.532	0.608	0.831	0.403	0.715	0.339	0.744	–	0.218	–	0.331	–	0.414	–	0.253
TaxoInstruct [70]	0.511	–	0.830	0.616	–	0.848	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
TEF [44]	0.635	0.728	0.861	0.643	0.734	0.881	0.541	0.625	0.828	–	–	–	–	–	–	–	–	–	–	–	–
FLAME [49]	0.634	–	0.851	0.632	–	0.825	0.587	–	0.781	0.472	0.742	–	–	–	–	–	–	–	–	–	–
CodeTaxo [99]	0.677	–	0.892	0.625	–	0.861	0.581	–	0.853	0.745	0.913	0.729	0.910	–	–	–	–	–	–	–	–
HyperExpan [37]	–	–	–	–	–	–	–	–	–	–	–	–	–	0.689	0.199	0.827	0.324	0.607	0.241	0.517	0.150
TaxoBell$_{BC}$ [48]	–	0.587	0.772	–	0.582	0.762	–	0.581	–	–	0.734	–	–	–	–	–	–	–	–	–	–
LORex [46]	0.673	–	0.829	0.647	–	0.874	0.553	–	0.843	0.495	0.845	–	–	–	–	–	–	–	–	–	–
Quantaxo [47]	0.492	0.595	0.791	0.573	0.674	0.821	–	–	–	0.258	0.710	–	–	–	–	–	–	–	–	–	–
TaxoLLaMA-bench [50]	–	–	–	–	–	–	–	–	–	–	–	–	–	0.302	–	0.314	–	0.459	–	0.519	–
Arborist [39]	–	0.337	0.525	–	0.412	0.612	–	0.453	–	–	0.533	–	–	0.602	–	0.722	–	0.435	–	0.280	–
Musubu [78]	0.453	–	0.654	0.449	–	0.762	0.423	–	0.724	0.285	0.640	0.354	0.752	–	–	–	–	–	–	–	–
DNG [102]	–	–	–	–	–	–	–	–	–	–	–	–	–	0.754	–	–	–	–	–	0.513	–

Wu & Palmer similarity are used to evaluate hierarchical consistency in embedding-based approaches [28]. Ranking-based variants such as Precision@K also appear in retrieval-style formulations [104]. More recent works, particularly LLM-based approaches, incorporate embedding-based evaluation (e.g., BERTScore in [81]) or qualitative and structural evaluation measures such as coherence and granularity [31]. Human or expert evaluation appears in a smaller subset of studies across different paradigms, including both classical and recent approaches, typically through manual relation validation, intrusion tests, or qualitative assessment of generated hierarchies [103]. However, such evaluation remains comparatively less common and less standardized than automatic metrics.

Overall, evaluation in taxonomy learning reflects the diversity of task formulations and methodological paradigms across the literature. Classification-oriented metrics such as precision, recall, F1-score, and accuracy remain standard in relation prediction and reconstruction settings, while ranking-based measures—including MRR, Hits@K, and mean rank—dominate expansion and completion tasks involving candidate selection. Although these metrics enable consistent comparison within specific experimental setups, cross-paper comparability remains challenging due to variations in datasets, candidate spaces, and evaluation protocols. These observations highlight the need for more standardized benchmarking practices and unified evaluation frameworks to support reproducible and meaningful comparison across taxonomy learning approaches.

C. Evaluation Protocols and Experimental Setups

Beyond dataset selection and metric choice, evaluation in taxonomy learning is strongly shaped by experimental protocols. A common setup is the held-out node protocol introduced above, which is widely used in expansion and completion tasks across SemEval, WordNet, and MAG benchmarks, including representative approaches such as TaxoExpan [68] and TMN [104]. While this setup enables controlled comparison, it often simplifies real-world taxonomy evolution scenarios.

Another important factor concerns candidate space design. Some studies evaluate models over the full taxonomy, whereas others restrict candidates through heuristic pruning or retrieval-based filtering strategies [28], [94]. These differences can substantially affect reported performance and complicate direct comparison across experimental settings.

Evaluation setups also vary in structural assumptions. Many approaches assume tree-structured taxonomies and primarily focus on leaf-node insertion (e.g., [36], [97]), whereas real-world taxonomies frequently exhibit DAG structures, multiple inheritance, and structural inconsistencies.

Finally, practical implementation details such as preprocessing pipelines, train–test split construction, and candidate generation strategies differ across studies, contributing to variation in reported results across otherwise similar benchmarks.

VI. CROSS-PARADIGM COMPARATIVE ANALYSIS

This section highlights key differences in supervision requirements, structural guarantees, scalability, interpretability, and robustness across the aforementioned methods, while also identifying complementary strengths among them.

A. Supervision and Learning Paradigms

Non-neural approaches are typically either unsupervised or rely on limited forms of supervision, such as seed patterns or manually defined rules. Pattern-based methods originating from early work such as Hearst-style extraction [23] rely on explicit linguistic cues rather than learned parameters, while rule-based systems depend on expert-designed heuristics and structural constraints. While these approaches offer high interpretability, their coverage and generalization are often limited to domains where such patterns or rules are applicable.

TABLE III

COMPARISON OF REPRESENTATIVE TAXONOMY COMPLETION METHODS ACROSS MULTIPLE BENCHMARKS. MOST RESULTS ARE TAKEN FROM THE ORIGINAL PAPERS; FOR METHODS WHERE SUITABLE RESULTS WERE NOT DIRECTLY AVAILABLE, VALUES ARE ADOPTED FROM PRIOR COMPARATIVE STUDIES. RESULTS MAY NOT BE DIRECTLY COMPARABLE DUE TO DIFFERENCES IN EXPERIMENTAL SETTINGS AND DATA SPLITS. "–" INDICATES THAT THE METRIC WAS NOT REPORTED OR NOT APPLICABLE. ↓ INDICATES THAT LOWER VALUES ARE BETTER.

Dataset	Metric	TMN [104]	GenTaxo‘ [100]	TaxoEnrich [27]	QEN [85]	TaxoComplete [5]	CoSTC [58]	COMI [94]	TacoPrompt [92]	ATTEMPT [89]	TAXBOX [95]	RAMA [101]
WordNet-Verb	MR↓	1445.801	2765.745	320.064	1802.404	589.3	241.089	109.454	436.799	–	1286	–
	MRR	0.304	0.428	0.452	0.340	–	0.505	0.615	0.557	0.330	–	–
	R@1	0.072	0.118	0.143	0.081	–	0.095	0.193	0.183	–	–	–
	R@5	0.163	0.208	0.252	0.186	–	0.278	0.399	0.369	–	–	–
	R@10	0.215	0.239	0.347	0.249	–	0.391	0.506	0.465	–	–	–
	P@1	0.108	0.235	0.276	0.124	–	–	–	–	–	0.179	–
	P@5	0.049	0.122	0.126	0.057	–	–	–	–	–	0.072	–
	P@10	0.032	0.066	0.081	0.038	–	–	–	–	–	0.045	–
	H@1	–	–	–	–	0.123	0.146	0.296	0.280	–	0.105	–
	H@5	–	–	–	–	0.316	0.392	0.555	0.523	–	0.212	–
	H@10	–	–	–	–	0.421	0.531	0.665	0.625	–	0.262	–
SemEval-Food	MR↓	–	–	–	353.733	–	61.471	25.321	47.423	–	281	–
	MRR	–	–	–	0.313	–	0.658	0.724	0.708	0.426	0.359	0.356
	R@1	–	–	–	0.070	–	0.187	0.289	0.309	–	–	–
	R@5	–	–	–	0.176	–	0.430	0.521	0.511	–	–	–
	R@10	–	–	–	0.234	–	0543	0.617	0.601	–	–	–
	P@1	–	–	–	0.146	–	–	–	–	–	0.318	–
	P@5	–	–	–	0.074	–	–	–	–	–	0.127	–
	P@10	–	–	–	0.049	–	–	–	–	–	0.071	–
	H@1	–	–	–	–	–	0.39	0.608	0.649	–	0.132	–
	H@5	–	–	–	–	–	0.734	0.878	0.858	–	0.264	–
	H@10	–	–	–	–	–	0.804	0.932	0.865	–	0.295	–
MAG-CS	MR↓	639.126	13213.731	87.789	–	1085.9	–	–	–	–	596	–
	MRR	0.204	0.239	0.578	–	–	–	–	–	–	0.240	–
	R@1	0.036	0.082	0.162	–	–	–	–	–	–	–	–
	R@5	0.099	0.185	0.434	–	–	–	–	–	–	–	–
	R@10	0.139	0.218	0.574	–	–	–	–	–	–	–	–
	P@1	0.156	0.254	0.274	–	–	–	–	–	–	0.238	–
	P@5	0.086	0.131	0.141	–	–	–	–	–	–	0.131	–
	P@10	0.060	0.085	0.093	–	–	–	–	–	–	0.087	–
	H@1	–	–	–	–	0.166	–	–	–	–	0.051	–
	H@5	–	–	–	–	0.346	–	–	–	–	0.139	–
	H@10	–	–	–	–	0.440	–	–	–	–	0.184	–
MAG-PSY	MR↓	212.298	7482.516	122.247	–	560.6	–	–	–	–	211	–
	MRR	0.471	0.464	0.583	–	–	–	–	–	–	0.479	–
	R@1	0.141	0.183	0.234	–	–	–	–	–	–	–	–
	R@5	0.305	0.402	0.424	–	–	–	–	–	–	–	–
	R@10	0.377	0.440	0.510	–	–	–	–	–	–	–	–
	P@1	0.287	0.376	0.374	–	–	–	–	–	–	0.328	–
	P@5	0.124	0.164	0.186	–	–	–	–	–	–	0.143	–
	P@10	0.077	0.090	0.124	–	–	–	–	–	–	0.089	–
	H@1	–	–	–	–	0.170	–	–	–	–	0.145	–
	H@5	–	–	–	–	0.392	–	–	–	–	0.317	–
	H@10	–	–	–	–	0.488	–	–	–	–	0.393	–
–	MR↓	–	–	–	344.735	–	109.081	29.477	49.140	–	–	–
MeSH	MRR	–	–	–	0.423	–	0.600	0.727	0.674	–	–	–
	R@1	–	–	–	0.071	–	0.110	0.199	0.179	–	–	–
	R@5	–	–	–	0.198	–	0.346	0.476	0.424	–	–	–
	R@10	–	–	–	0.165	–	0.475	0.615	0.559	–	–	–
	P@1	–	–	–	0.091	–	–	–	–	–	–	–
	P@5	–	–	–	0.066	–	–	–	–	–	–	–
	P@10	–	–	–	–	–	–	–	–	–	–	–
	H@1	–	–	–	–	–	0.249	0.453	0.407	–	–	–
	H@5	–	–	–	–	–	0.615	0.794	0.746	–	–	–
	H@10	–	–	–	–	–	0.726	0.885	84.6	–	–	–

In contrast, neural representation learning approaches introduce supervised, weakly supervised, or self-supervised training regimes. Embedding-based models, including hyperbolic representations [37] and box embeddings [28], learn latent hierarchical structure from data. Self-supervised expansion frameworks such as STEAM [97] and contrastive learning approaches [58] reduce reliance on manual labels but still require large corpora or structured seed taxonomies. LLM-based approaches further shift this paradigm by leveraging large pre-trained generative models [77]. Many systems rely on prompt-based or in-context learning rather than task-specific supervised training [91], [98]. While this reduces explicit annotation requirements, supervision becomes implicit in the pretraining data, raising concerns regarding controllability, bias, and reproducibility.

In practice, these paradigms differ not only in explicit supervision but also in their reliance on external resources, with neural and LLM-based methods typically depending more heavily on pretrained corpora and representations.

TABLE IV
COMPARISON OF TAXONOMY CONSTRUCTION METHODS ACROSS DATASETS. METRICS ARE EXPLICITLY DISTINGUISHED AS EDGE-BASED OR ANCESTOR-BASED WHEN APPLICABLE. RESULTS ARE TAKEN FROM ORIGINAL PAPERS AND MAY NOT BE DIRECTLY COMPARABLE DUE TO DIFFERENCES IN SETTINGS. "–" INDICATES THAT THE METRIC WAS NOT REPORTED OR NOT APPLICABLE. WIKI REFERS TO SUBSETS OF ENGLISH WIKIPEDIA CORPORA AS USED IN PRIOR WORK.

Dataset	Metric	CTP [10]	HiExpan' [69]	Graph2Taxo [66]	DTaxa [22]	CoL [98]	Pietrasik et al [63]	Jain & Anke [26]
WordNet	Ancestor-P	0.693	–	–	–	–	–	–
	Ancestor-R	0.662	–	–	–	–	–	–
	Ancestor-F1	0.667	–	–	–	–	–	–
	Edge-F1	–	–	–	–	–	0.550	–
SemEval-Sci	Ancestor-P	–	–	–	–	0.912	–	–
	Ancestor-R	–	–	–	–	0.481	–	–
	Ancestor-F1	–	–	–	–	0.626	–	–
	Edge-P	0.294	–	0.82	0.597	0.596	–	0.393
	Edge-R	0.288	–	0.33	0.314	0.460	–	0.367
	Edge-F1	0.291	–	0.47	0.403	0.515	–	0.379
DBLP	Ancestor-P	–	0.843	–	–	0.799	–	–
	Ancestor-R	–	0.376	–	–	0.630	–	–
	Ancestor-F1	–	0.520	–	–	0.688	–	–
	Edge-P	–	0.829	–	–	0.550	–	–
	Edge-R	–	0.460	–	–	0.442	–	–
	Edge-F1	–	0.592	–	–	0.479	–	–
Wiki	Ancestor-P	–	0.847	–	–	0.991	–	–
	Ancestor-R	–	0.725	–	–	0.959	–	–
	Ancestor-F1	–	0.781	–	–	0.975	–	–
	Edge-P	–	0.848	–	–	0.979	–	–
	Edge-R	–	0.702	–	–	0.949	–	–
	Edge-F1	–	0.768	–	–	0.964	–	–

B. Structural Guarantees and Hierarchical Consistency

Taxonomies are inherently structured objects requiring acyclicity, transitivity, and hierarchical consistency. Non-neural methods often enforce such constraints explicitly through rule-based systems or graph-based optimization. In contrast, neural embedding approaches attempt to encode hierarchy geometrically, for example through hyperbolic space representations [37] or box-based containment modeling [105]. While these models introduce structural inductive biases, they may still produce inconsistencies without explicit constraint enforcement or post-processing. LLM-based approaches face greater structural challenges. Generative models may produce plausible yet structurally inconsistent hierarchies, as structural constraints are not inherently enforced during generation. Recent works attempt to address this through iterative refinement or multi-stage prompting [15], [98]. However, these stepwise procedures mainly validate local decisions and do not necessarily guarantee global structural validity across the entire hierarchy, where errors can accumulate or interact across branches.

C. Scalability and Computational Cost

Symbolic and pattern-based systems scale relatively well when extraction rules are simple, but they often struggle with sparse linguistic evidence and domain adaptation. Neural models, by comparison, scale effectively with data and computational resources, particularly when trained on large corpora. However, training complex embedding or graph-based models can be computationally intensive. LLM-based approaches benefit from powerful pretrained models but incur high inference costs due to large context sizes and repeated querying, along with limited transparency regarding internal representations. Additionally, dependence on proprietary APIs in some cases raises reproducibility and cost concerns.

More broadly, scalability is not solely determined by model complexity but also by candidate search space design and inference strategy.

D. Interpretability and Transparency

Interpretability differs substantially across paradigms. Rule-based and pattern-based methods are inherently interpretable, as extracted relations can be traced to explicit linguistic patterns. Neural representation learning methods provide partial interpretability through geometric structure (e.g., hierarchical distance in hyperbolic space), but these representations remain largely latent and difficult to interpret directly. LLM-based systems tend to be the least transparent, as hierarchical decisions are embedded within large-scale generative processes. Although prompt design provides some degree of control, the underlying reasoning does not necessarily correspond to explicit structural guarantees.

E. Robustness and Generalization

Non-neural systems, while stable and deterministic, typically require manual adaptation to new domains and struggle in noisy or sparse settings. Neural and embedding-based models often generalize better in such conditions, particularly when trained with contrastive or self-supervised objectives [58]. LLM-based methods demonstrate strong zero-shot generalization capabilities [77], but may suffer from hallucination, structural inconsistency, and sensitivity to prompt formulation.

These differences suggest that robustness remains closely tied to both training data diversity and the degree of structural inductive bias imposed by each paradigm.

F. Synthesis

Overall, taxonomy learning research reflects a clear progression from explicit symbolic reasoning toward representation learning and, more recently, generative reasoning with LLMs. Each paradigm introduces distinct trade-offs across interpretability, scalability, structural guarantees, and flexibility. Non-neural approaches emphasize control and transparency but face limitations in scalability and coverage. Neural representation learning methods offer scalable modeling with partial structural inductive biases but remain sensitive to training data and representation quality. LLM-based approaches provide flexible reasoning and strong zero-shot capabilities, yet introduce challenges related to structural consistency, hallucination, and reproducibility.

Taken together, these observations suggest that future progress may emerge from hybrid frameworks that combine symbolic constraints, neural representations, and generative reasoning to balance flexibility with structural reliability.

VII. Challenges and Future Directions

Despite substantial methodological progress, taxonomy learning remains characterized by persistent challenges spanning structural consistency, evaluation standardization, scalability, and paradigm integration. Drawing from the limitations reported across the curated literature, we summarize key open problems and outline promising research directions.

A. Structural Validity and Global Consistency

Ensuring hierarchical consistency remains a central challenge. Taxonomies require properties such as acyclicity, transitivity, and coherent depth organization. While symbolic systems explicitly enforce structural constraints, neural and embedding-based models may produce locally accurate but globally inconsistent hierarchies, particularly due to the difficulty of modeling asymmetric and directional inheritance in hierarchical relations. Hyperbolic and box-based representations attempt to encode structural bias geometrically [37], [105]. Nevertheless, violations of structural constraints, such as inconsistent parent–child relations or failure to preserve transitivity, can still occur.

LLM-based approaches further amplify this issue. Generative models may produce plausible but structurally invalid relations, particularly in zero-shot or prompt-driven settings [77], [98]. Designing mechanisms that combine generative flexibility with strict structural guarantees remains an open research problem.

B. Evaluation Fragmentation and Benchmark Standardization

As discussed in Section V, evaluation practices remain fragmented and are often based on artificial held-out node protocols that may not fully reflect real-world taxonomy evolution. A limited number of benchmarks, most notably WordNet subsets and SemEval-2016 Task 14, dominate empirical evaluation. However, dataset splits, preprocessing pipelines, and task formulations often differ across studies, limiting reproducibility and cross-paradigm comparison.

Furthermore, LLM-based works increasingly incorporate human evaluation or qualitative assessment, introducing additional variability. Establishing standardized benchmark splits, hierarchy-aware metrics, and unified evaluation protocols is essential for meaningful comparison across symbolic, neural, and LLM-based paradigms.

A particularly underexplored challenge arises in taxonomy refinement, where the lack of shared baselines and controlled evaluation settings makes it difficult to compare methods and isolate the impact of different refinement strategies.

Future work could address this limitation by developing dedicated refinement benchmarks, for example by constructing taxonomies with controlled perturbations (e.g., injected noise) and evaluating methods based on their ability to recover the original structure. Such standardized setups would enable more consistent and comparable evaluation across approaches.

C. Data Dependence and Domain Adaptation

Many methods rely heavily on domain-specific corpora or curated seed taxonomies. Pattern-based systems require explicit linguistic cues, neural models depend on large corpora for representation learning, and LLM-based approaches rely on implicit knowledge encoded during large-scale pretraining. Cross-domain transfer remains challenging, particularly for specialized or low-resource domains.

Recent work on self-supervised and low-resource taxonomy expansion [49], [78] suggests promising directions, yet robust cross-domain generalization is not fully solved.

D. Scalability and Computational Cost

While neural and LLM-based systems scale well in representational capacity, they often incur significant computational cost. Training graph neural networks or contrastive models can be resource-intensive, and LLM-based approaches may depend on expensive inference pipelines or proprietary APIs. Balancing scalability with reproducibility and cost efficiency remains a practical concern.

E. Interpretability and Human-in-the-Loop Systems

Interpretability remains uneven across paradigms. Rule-based systems are transparent but limited in flexibility, whereas neural and LLM-based models offer improved performance at the cost of explainability.

Several recent works explore interactive or semi-automatic refinement settings [51], [61], incorporating expert feedback or iterative correction mechanisms. Human-in-the-loop frameworks may offer a pragmatic compromise between structural reliability and automation.

F. Toward Hybrid and Unified Frameworks

The comparative analysis suggests that no single paradigm fully addresses the requirements of taxonomy learning. Symbolic approaches provide structural control, neural embeddings capture scalable semantic representation, and LLMs introduce flexible generative reasoning.

Future research may benefit from hybrid frameworks that integrate explicit structural constraints, learned hierarchical embeddings, and controlled generative prompting. Such integration could enable scalable, interpretable, and structurally valid taxonomy construction across domains.

VIII. Conclusion

This survey has presented a comprehensive methodological analysis of taxonomy learning research, spanning construction, expansion, completion, and refinement tasks. Rather than organizing prior work solely by downstream tasks, we introduced a method-driven taxonomy that classifies approaches into non-neural, neural representation learning, and LLM–based methods.

Through this structured perspective, we examined the modeling assumptions, supervision strategies, structural properties, and evaluation practices characterizing each paradigm. Our quantitative synthesis further highlighted the concentration of benchmarks around WordNet and SemEval resources, the dominance of classification and ranking metrics, and the fragmentation of evaluation protocols across paradigms.

The comparative analysis reveals that each methodological family offers complementary strengths: symbolic approaches provide interpretability and explicit structural control; neural embedding models enable scalable semantic representation; and LLM-based systems introduce flexible generative reasoning. However, challenges remain in ensuring structural validity, establishing standardized evaluation frameworks, and enabling reproducible cross-paradigm comparison.

By consolidating methodological trends and empirical practices under a unified framework, this survey aims to serve as a reference point for future research in taxonomy learning. We envision that the proposed paradigm-based classification will facilitate clearer comparison, encourage hybrid modeling strategies, and contribute to the development of more robust, scalable, and structurally consistent taxonomy systems.

References

[1] Rami Aly, Shantanu Acharya, Alexander Ossa, Arne Köhn, Chris Biemann, and Alexander Panchenko. Every child should have parents: a taxonomy refinement algorithm based on hyperbolic term embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[2] H Angermann. Taxomulti: Rule-based expert system to customize product taxonomies for multi-channel e-commerce. sn computer science, vol. 3, article 177, 2022.

[3] Luis Espinosa Anke, Jose Camacho-Collados, Sara Rodríguez-Fernández, Horacio Saggion, and Leo Wanner. Extending wordnet with fine-grained collocational information via supervised distributional learning. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: technical papers*, 2016.

[4] Luis Espinosa Anke, Francesco Ronzano, and Horacio Saggion. Taln at semeval-2016 task 14: Semantic taxonomy enrichment via sense-based embeddings. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016.

[5] Ines Arous, Ljiljana Dolamic, and Philippe Cudré-Mauroux. Taxocomplete: Self-supervised taxonomy completion leveraging position-enhanced semantic matching. In *Proceedings of the ACM Web Conference 2023*, 2023.

[6] Muhammad Arslan and Christophe Cruz. Semantic taxonomy enrichment to improve business text classification for dynamic environments. In *2022 international conference on innovations in intelligent systems and applications (INISTA)*. IEEE, 2022.

[7] Mohit Bansal, David Burkett, Gerard De Melo, and Dan Klein. Structured learning for taxonomy induction with belief propagation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

[8] Georgeta Bordea, Els Lefever, and Paul Buitelaar. Semeval-2016 task 13: Taxonomy extraction evaluation (texeval-2). In *Proceedings of the 10th international workshop on semantic evaluation (semeval-2016)*, 2016.

[9] Sabur Butt, Gustavo De los Ríos Alatorre, Luis José González Gómez, and Hector G Ceballos. Semi-supervised taxonomy expansion and completion in dynamic taxonomies. *Applied Sciences*, 15(12):6517, 2025.

[10] Catherine Chen, Kevin Lin, and Dan Klein. Constructing taxonomies from pretrained language models. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, 2021.

[11] Derek Chen, Howard Chen, Yi Yang, Alexander Lin, and Zhou Yu. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021.

[12] Xiaoli Chen. Evolving taxonomy based on graph neural networks. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 2020.

[13] Ying Chen, Lianming Zhang, Jiusheng Li, and Pingping Dong. Hyperbolic graph representation learning: methods, applications and challenges—a survey. *Neurocomputing*, page 131044, 2025.

[14] Sijie Cheng, Zhouhong Gu, Bang Liu, Rui Xie, Wei Wu, and Yanghua Xiao. Learning what you need from what you did: Product taxonomy expansion with user behaviors supervision. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, 2022.

[15] Simone D'Amico, Alessia De Santo, Mario Mezzanzanica, and Fabio Mercorio. Taxonomy expansion through collaborative llm mapping. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, 2025.

[16] Simone D'Amico, Alessia De Santo, Fabio Mercorio, and Mario Mezzanzanica. Enriching skill taxonomies through vector space models. In *2024 IEEE International Conference on Big Data (BigData)*, 2024.

[17] Christiane Fellbaum. Wordnet. *WordNet An Electronic Lexical Database*, page 69, 1998.

[18] Genet Asefa Gesese, Russa Biswas, Mehwish Alam, and Harald Sack. A survey on knowledge graph embeddings with literals: Which model links better literal-ly? *Semantic Web*, 12(4), 2021.

[19] Zeinab Ghamlouch and Mehwish Alam. Enriching taxonomies using large language models. In *ECAI 2025-28th European Conference on Artificial Intelligence (Demo Track)*, 2025.

[20] Anna Giabelli, Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Andrea Seveso. Neo: A tool for taxonomy enrichment with new emerging occupations. In *International Semantic Web Conference*, pages 568–584. Springer, 2020.

[21] Amit Gupta, Rémi Lebret, Hamza Harkous, and Karl Aberer. Taxonomy induction using hypernym subsequences. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017.

[22] Yongming Han, Yanwei Lang, Minjie Cheng, Zhiqiang Geng, Guofei Chen, and Tao Xia. Dtaxa: An actor–critic for automatic taxonomy induction. *Engineering Applications of Artificial Intelligence*, 106:104501, 2021.

[23] M Hearst. Automatic acquisition of hyponyms from large text corpora in proc. In *14th international conference computational linguistics*, 1992.

[24] Yujia Hu, Tuan-Phong Nguyen, Shrestha Ghosh, and Simon Razniewski. Enabling llm knowledge analysis via extensive materialization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

[25] Jiaxin Huang, Yiqing Xie, Yu Meng, Yunyi Zhang, and Jiawei Han. Corel: Seed-guided topical taxonomy construction by concept learning and relation transferring. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining*, 2020.

[26] Devansh Jain and Luis Espinosa Anke. Distilling hypernymy relations from language models: On the effectiveness of zero-shot taxonomy induction. In *Proceedings of the 11th joint conference on lexical and computational semantics*, 2022.

[27] Minhao Jiang, Xiangchen Song, Jieyu Zhang, and Jiawei Han. Taxoenrich: Self-supervised taxonomy completion via structure-semantic representations. In *Proceedings of the ACM web conference 2022*, 2022.

[28] Song Jiang, Qiyue Yao, Qifan Wang, and Yizhou Sun. A single vector is not enough: Taxonomy expansion via box embeddings. In *Proceedings of the ACM web conference 2023*, 2023.

[29] David Jurgens and Mohammad Taher Pilehvar. Reserating the awesometastic: An automatic extension of the wordnet taxonomy for novel terms. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015.

[30] David Jurgens and Mohammad Taher Pilehvar. Semeval-2016 task 14: Semantic taxonomy enrichment. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016.

[31] Priyanka Kargupta, Nan Zhang, Yunyi Zhang, Rui Zhang, Prasenjit Mitra, and Jiawei Han. Taxoadapt: Aligning llm-based multidimensional taxonomy construction to evolving research corpora. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025.

[32] Zornitsa Kozareva and Eduard Hovy. A semi-supervised method to learn and construct taxonomies using the web. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, 2010.

[33] Dongha Lee, Jiaming Shen, SeongKu Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters. In *Proceedings of the ACM Web Conference 2022*, 2022.

[34] Dongha Lee, Jiaming Shen, Seonghyeon Lee, Susik Yoon, Hwanjo Yu, and Jiawei Han. Topic taxonomy expansion via hierarchy-aware topic phrase generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1687–1700, 2022.

[35] Zequn Liu, Shukai Wang, Yiyang Gu, Ruiyi Zhang, Ming Zhang, and Sheng Wang. Graphine: A dataset for graph-aware terminology definition generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[36] Zichen Liu, Hongyuan Xu, Yanlong Wen, Ning Jiang, Haiying Wu, and Xiaojie Yuan. Temp: Taxonomy expansion with dynamic margin loss through taxonomy-paths. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, 2021.

[37] Mingyu Derek Ma, Muhao Chen, Te-Lin Wu, and Nanyun Peng. Hyperexpan: Taxonomy expansion with hyperbolic representation learning. In *Findings of the Association for Computational Linguistics: EMNLP*, 2021.

[38] Lorenzo Malandri, Fabio Mercorio, Mario Mezzanzanica, and Navid Nobani. Taxoref: Embeddings evaluation for ai-driven taxonomy refinement. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2021.

[39] Emaad Manzoor, Rui Li, Dhananjay Shrouty, and Jure Leskovec. Expanding taxonomies with implicit edge semantics. In *Proceedings of the web conference 2020*, pages 2044–2054, 2020.

[40] Yuning Mao, Tong Zhao, Andrey Kan, Chenwei Zhang, Xin Luna Dong, Christos Faloutsos, and Jiawei Han. Octet: Online catalog taxonomy enrichment with self-supervision. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 2020.

[41] Daniele Margiotta, Danilo Croce, and Roberto Basili. Taxosbert: Unsupervised taxonomy expansion through expressive semantic similarity. In *International Conference on Deep Learning Theory and Applications*, 2023.

[42] Félix Martel and Amal Zouaq. Taxonomy extraction using knowledge graph embeddings and hierarchical clustering. In *Proceedings of the 36th Annual ACM Symposium on applied computing*, 2021.

[43] Yuan Meng, Songlin Zhai, Zhihua Chai, Yuxin Zhang, Tianxing Wu, Guilin Qi, and Wei Song. Which is better? taxonomy induction with learning the optimal structure via contrastive learning. *Knowledge-Based Systems*, 304:112405, 2024.

[44] Yuan Meng, Songlin Zhai, Yuxin Zhang, Zhongjian Hu, and Guilin Qi. Tef: Causality-aware taxonomy expansion via front-door criterion. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2025.

[45] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[46] Sahil Mishra, Kumar Arjun, and Tanmoy Chakraborty. Rank, chunk and expand: Lineage-oriented reasoning for taxonomy expansion. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

[47] Sahil Mishra, Avi Patni, Niladri Chatterjee, and Tanmoy Chakraborty. Quantaxo: A quantum approach to self-supervised taxonomy expansion. *arXiv preprint arXiv:2501.14011*, 2025.

[48] Sahil Mishra, Srinitish Srinivasan, Srikanta Bedathur, and Tanmoy Chakraborty. Taxobell: Gaussian box embeddings for self-supervised taxonomy expansion. *arXiv preprint arXiv:2601.09633*, 2026.

[49] Sahil Mishra, Ujjwal Sudev, and Tanmoy Chakraborty. Flame: Self-supervised low-resource taxonomy expansion using large language models. *ACM Transactions on Intelligent Systems and Technology*, 2024.

[50] Viktor Moskvoretskii, Ekaterina Neminova, Alina Lobanova, Alexander Panchenko, and Irina Nikishina. Taxollama: Wordnet-based model for solving multiple lexical semantic tasks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

[51] Elham Motamedi, Inna Novalija, and Luis Rei. Semi-automatic hierarchical taxonomy creation from existing taxonomies with large language models. *Business & Information Systems Engineering*, 2026.

[52] Maryam Mousavi, Elena Steiner, Steven R. Corman, Scott Ruston, Dylan Weber, and Hasan Davulcu. Stif: Semi-supervised taxonomy induction using term embeddings and clustering. In *Proceedings of the 2021 5th International Conference on Natural Language Processing and Information Retrieval*, 2021.

[53] Azad Naik and Huzefa Rangwala. Filter based taxonomy modification for improving hierarchical classification. *arXiv preprint arXiv:1603.00772*, 2016.

[54] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010.

[55] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

[56] Irina Nikishina, Varvara Logacheva, Alexander Panchenko, and Natalia Loukachevitch. Studying taxonomy enrichment on diachronic wordnet versions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3095–3106, 2020.

[57] Kiyoshi Nitta. Improving taxonomies for large-scale hierarchical classifiers of web documents. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010.

[58] Yuhang Niu, Hongyuan Xu, Ciyi Liu, Yanlong Wen, and Xiaojie Yuan. Contrastive representation learning for self-supervised taxonomy completion. In *IJCAI*, volume 8, pages 6442–6450, 2024.

[59] Alexander Panchenko, Stefano Faralli, Eugen Ruppert, Steffen Remus, Hubert Naets, Cédrick Fairon, Simone Paolo Ponzetto, and Chris Biemann. Taxi at semeval-2016 task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016.

[60] Heiko Paulheim. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*, 8(3), 2017.

[61] Yiwen Peng, Thomas Bonald, and Mehwish Alam. Refining wikidata taxonomy using large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 5395–5399, 2024.

[62] Bornali Phukon, Anasua Mitra, Ranbir Sanasam, and Priyankoo Sarmah. Team: A multitask learning based taxonomy expansion approach for attach and merge. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 2022.

[63] Marcin Pietrasik, Marek Reformat, and Anna Wilbik. Non-parametric path based model for taxonomy induction in knowledge graphs. In *Belgium netherlands conference on artificial intelligence*, 2024.

[64] Jiaming Qu, Madhu Gopinathan, Shayan Ali Akbar, and Omar Alonso. Interactive taxonomy development with hybrid methods. 2026.

[65] Heidi Sand, Erik Velldal, and Lilja Øvrelid. Wordnet extension via word embeddings: Experiments on the norwegian wordnet. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 298–302, 2017.

[66] Chao Shang, Sarthak Dash, Md Faisal Mahbub Chowdhury, Nandana Mihindukulasooriya, and Alfio Gliozzo. Taxonomy construction of unseen domains via graph-based cross-domain knowledge transfer. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, 2020.

[67] Jingbo Shang, Xinyang Zhang, Liyuan Liu, Sha Li, and Jiawei Han. Nettaxo: Automated topic taxonomy construction from text-rich network. In *Proceedings of the web conference 2020*, 2020.

[68] Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. Taxoexpan: Self-supervised taxonomy expansion with position-enhanced graph neural network. In *Proceedings of the web conference 2020*, 2020.

[69] Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T Vanni, Brian M Sadler, and Jiawei Han. Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018.

[70] Yanzhen Shen, Yu Zhang, Yunyi Zhang, and Jiawei Han. A unified taxonomy-guided instruction tuning framework for entity set expansion and taxonomy expansion. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3208–3220, 2025.

[71] Jingchuan Shi, Hang Dong, Jiaoyan Chen, Zhe Wu, and Ian Horrocks. Taxonomy completion via implicit concept insertion. In *Proceedings of the ACM Web Conference 2024*, 2024.

[72] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, 2015.

[73] Rion Snow, Dan Jurafsky, and Andrew Y Ng. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics*, 2006.

[74] Arnolnt Spyros, Anna Kougioumtzidou, Angelos Papoutsis, Eleni Darra, Dimitrios Kavallieros, Athanasios Tziouvaras, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris. A comprehensive survey of manual and dynamic approaches for cybersecurity taxonomy generation. *Knowledge and Information Systems*, 67, 2025.

[75] Fabian M Suchanek, Mehwish Alam, Thomas Bonald, Lihu Chen, Pierre-Henri Paris, and Jules Soria. Yago 4.5: A large and clean knowledge base with a rich taxonomy. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 131–140, 2024.

[76] Kai Sun, Jifan Yu, Juanzi Li, and Lei Hou. Exploring sequence-to-sequence taxonomy expansion via language model probing. *Expert Systems with Applications*, 239:122321, 2024.

[77] Yushi Sun, Hao Xin, Kai Sun, Yifan Ethan Xu, Xiao Yang, Xin Luna Dong, Nan Tang, and Lei Chen. Are large language models a good replacement of taxonomies? *arXiv preprint arXiv:2406.11131*, 2024.

[78] Kunihiro Takeoka, Kosuke Akimoto, and Masafumi Oyamada. Low-resource taxonomy enrichment with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[79] Hristo Tanev and Agata Rotondi. Deftor at semeval-2016 task 14: Taxonomy enrichment using definition vectors. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016.

[80] Nikhita Vedula, Patrick K Nicholson, Deepak Ajwani, Sourav Dutta, Alessandra Sala, and Srinivasan Parthasarathy. Enriching taxonomies with functional domain knowledge. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018.

[81] Binh Vu, Rashmi Govindraju Naik, Bao Khanh Nguyen, Sina Mehraeen, and Matthias Hemmje. Automated taxonomy construction using large language models: A comparative study of fine-tuning and prompt engineering. *Eng*, 6(11):283, 2025.

[82] Chengyu Wang, Xiaofeng He, and Aoying Zhou. A short survey on taxonomy learning from text corpora: Issues, resources and recent advances. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.

[83] Jingjing Wang, Changsung Kang, Yi Chang, and Jiawei Han. A hierarchical dirichlet model for taxonomy expansion for search engines. In *Proceedings of the 23rd international conference on World wide web*, 2014.

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

[85] Suyuchen Wang, Ruihui Zhao, Yefeng Zheng, and Bang Liu. Qen: Applicable taxonomy completion via evaluating full taxonomic relations. In *Proceedings of the ACM Web Conference 2022*, 2022.

[86] Xin Wang, Zirui Chen, Haofen Wang, Leong Hou U, Zhao Li, and Wenbin Guo. Large language model enhanced knowledge representation learning: A survey. *Data Science and Engineering*, 2025.

[87] Yuquan Wang, Yanpeng Wang, Yiming Mao, Jifan Yu, Kaisheng Zeng, Lei Hou, Juanzi Li, and Jie Tang. Expertise-aware crowdsourcing taxonomy enrichment. In *International Conference on Web Information Systems Engineering*. Springer, 2021.

[88] Minxin Wu, Yifei Gong, Heping Lu, Baofeng Li, Kai Wang, Yanquan Zhou, and Lei Li. Large models and multimodal: A survey of cutting-edge approaches to knowledge graph completion. *Data Science and Engineering*, 10, 2025.

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

[90] Fred Xu, Song Jiang, Zijie Huang, Xiao Luo, Shichang Zhang, Yuanzhou Chen, and Yizhou Sun. Fuse: Measure-theoretic compact fuzzy set representation for taxonomy expansion. In *Findings of the association for computational linguistics: ACL 2024*, pages 2707–2720, 2024.

[91] Hongyuan Xu, Yunong Chen, Zichen Liu, Yanlong Wen, and Xiaojie Yuan. Taxoprompt: A prompt-based generation method with taxonomic context for self-supervised taxonomy expansion. In *IJCAI*, volume 22, pages 4432–4438, 2022.

[92] Hongyuan Xu, Ciyi Liu, Yuhang Niu, Yunong Chen, Xiangrui Cai, Yanlong Wen, and Xiaojie Yuan. Tacoprompt: A collaborative multi-task prompt learning method for self-supervised taxonomy completion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[93] Hongyuan Xu, Yuhang Niu, Ciyi Liu, Yanlong Wen, and Xiaojie Yuan. Taxopro: A plug-in lora-based cross-domain method for low-resource taxonomy completion. *Transactions of the Association for Computational Linguistics*, 13, 2025.

[94] Hongyuan Xu, Yuhang Niu, Yanlong Wen, and Xiaojie Yuan. Compress and mix: Advancing efficient taxonomy completion with large language models. In *Proceedings of the ACM on Web Conference 2025*, 2025.

[95] Wei Xue, Yongliang Shen, Wenqi Ren, Jietian Guo, Shiliang Pu, and Weiming Lu. Insert or attach: Taxonomy completion via box embedding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

[96] Hui Yang. Constructing task-specific taxonomies for document collection browsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012.

[97] Yue Yu, Yinghao Li, Jiaming Shen, Hao Feng, Jimeng Sun, and Chao Zhang. Steam: Self-supervised taxonomy expansion with mini-paths. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1026–1035, 2020.

[98] Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Shangbin Feng, Zhenwen Liang, Zhihan Zhang, and Meng Jiang. Chain-of-layer: Iteratively prompting large language models for taxonomy induction from limited examples. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2024.

[99] Qingkai Zeng, Yuyang Bai, Zhaoxuan Tan, Zhenyu Wu, Shangbin Feng, and Meng Jiang. Codetaxo: Enhancing taxonomy expansion with limited examples via code language prompts. In *Findings of the Association for Computational Linguistics: ACL*, 2025.

[100] Qingkai Zeng, Jinfeng Lin, Wenhao Yu, Jane Cleland-Huang, and Meng Jiang. Enhancing taxonomy completion with concept generation via fusing relational representations. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021.

[101] Qingkai Zeng, Zhihan Zhang, Jinfeng Lin, and Meng Jiang. Completing taxonomies with relation-aware mutual attentions. In *KDD*, 2023.

[102] Songlin Zhai, Weiqing Wang, Yuanfang Li, and Yuan Meng. Dng: taxonomy expansion by exploring the intrinsic directed structure on non-gaussian space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2023.

[103] Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. Taxogen: Constructing topical concept taxonomy by adaptive term embedding and clustering. *Proc. KDDI*, 2018.

[104] Jieyu Zhang, Xiangchen Song, Ying Zeng, Jiaze Chen, Jiaming Shen, Yuning Mao, and Lei Li. Taxonomy completion via triplet matching network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 2021.

[105] Yuhang Zhang, Jiwei Qin, and Chongren Feng. Peb-taxo: Projecting entities as boxes for taxonomy expansion. *Neural Processing Letters*, 56(2):102, 2024.

[106] Tinghui Zhu, Jingping Liu, Jiaqing Liang, Haiyun Jiang, Yanghua Xiao, Zongyu Wang, Rui Xie, and Yunsen Xian. Towards visual taxonomy expansion. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023.

TABLE V
TASK × PARADIGM MATRIX OF THE PAPERS INCLUDED IN THIS SURVEY, ILLUSTRATING THEIR DISTRIBUTION ACROSS TAXONOMY LEARNING TASKS AND MODELING PARADIGMS.

Learning task	Non-neural	Neural	LLM-based
Construction	Marti A. Hearst [23] Hui Yang [96] Gupta et al. [21] Pietrasik et al. [63]	Meng et al. [43] **CTP** [10] Martel & Zouaq [42] **HiExpan** [69] Jain & Anke [26] **Graph2Taxo** [66] **STIF** [52] **NetTaxo** [67] **CoRel** [25] **DTaxa** [22] **TaxoGen** [103]	**TaxoLLaMA** [50] **TaxoAdapt** [31] Vu et al. [81] Hu et al. [24] **Chain-of-Layer** [98]
Expansion	**CROWN** [29] Wang et al. [83] Wang et al. [87] Snow et al. [73]	**OCTET** [40] **TaxoExpan** [68] **PEB-TAXO** [105] **VTE** [106] **TaxoSBERT** [41] **BoxTaxo** [28] Cheng et al. [14] **Musubu** [78] **HyperExpan** [37] **ColWordNet (CWN)** [3] **ETF** [80] **NEO** [20] **Arborist** [39] Nikishina et al. [56] **STEAM** [97] **TaxoBell** [48] **TEMP** [36] **FUSE** [90] **HEF** [84] **LORex** [46] **TaxoSeq** [76] **QuantaXO** [47] **DNG** [102] **TEF** [44] **TEAM** [62] Xiaoli Chen [12] Sand et al. [65] **TopicExpan** [34] Arslan & Cruz [6] Butt et al. [9]	**FLAME** [49] **TAXMAP** [15] **Taxoria** [19] **CodeTaxo** [99] **TAXOINSTRUCT** [70] **TaxoPrompt** [91] Qu et al. [64] **TaxoLLaMA** [50]
Completion	**Deftor** [79]	**TaxoEnrich** [27] **GenTaxo** [100] **TaxoPro** [93] Butt et al. [9] **CoSTC** [58] **ICON** [71] **TaxoComplete** [5] **TaxoCom** [33] **TMN** [104] **TacoPrompt** [92] **QEN** [85] **ATTEMPT** [89] **TAXBOX** [95] **RAMA** [101] Anke et al. [4]	**COMI** [94]
Refinement	**rewHier** [53] Kiyoshi Nitta [57] **TaxoMulti** [2] **YAGO 4.5** [75]	**TaxoRef** [38] Aly et al. [1]	**WiKC** [61] Motamedi et al. [51]