Definition grounding for Taxonomy construction

Brice Tsakam, brice_tsakam@hotmail.com
Horgen Switzerland, April 2026

Abstract

Taxonomy construction consists in building a tree of parent-child entities, specifically, hypernym hyponym pairs. When tackling Natural Language Processing (NLP) tasks with a Large Language Model (LLM), we may rely on the broad model's knowledge derived from its training data, typically a large text collection obtained from public internet crawls. However the understanding of relationships between objects and concepts in a domain specific setting often differ from common knowledge or common sense. Using small scale experiments, we show that defining each taxon in context makes constructing specialized zoological taxonomies reliable.

Introduction

Taxonomy construction is an essential knowledge management with applications in knowledge extraction and search, among others. LLMs trained on crawls of internet data, including wikipedia hold broad common knowledge, including biological taxonomy knowledge[1]. We choose to limit ourselves to a small subset of the Zoological Nomenclature[1] because it is well defined and stable. It is largely included in the training data of modern LLMs[11].

The paper is organized as follows. Section 1 presents the taxonomy construction method. Section 2 evaluates and compares two approaches to effectively test the ability of the model to make use of its parametric taxonomy knowledge by conducting taxonomy reconstruction under two settings. In the first setting, we rely on a pre-trained model's knowledge of the zoological entities (without additional definition), i.e. parametric memory, to identify parent-child relationships. In the second setting, we leverage in context learning by introducing each taxon's definition. Section 3 discusses related research before a concluding remark.

Section 1: Taxonomy construction method

The taxonomy construction proceeds in 2 steps.

Step1: Detection of parent-child relationships using machine reading comprehension. We basically submit pairs of entities' names and prompt the LLM to decide whether or not they have a parent-child relationship.

Step2: Building the taxonomy tree out of the connected entities. The connected entities make up a directed forest. We proceed by pruning redundant relationships. The taxonomy tree is obtained as the transitive reduction[3] of the directed forest -

which works providing the graph made of all identified parent-child relationships from **Step 1** is acyclic and fully connected.

This 2 steps sequence is subject to the hard requirement that for **Step 2** to proceed, **Step 1** must return an acyclic graph.

Parent-Child detection

We present 2 approaches for the parent-child relationship decision in **Step 1**. With both parent-child detection methods we keep **Step 2** unchanged.

Entity name only: without definition

We tackle **Step 1**, the parent-child relationship detection, using machine reading comprehension. We assume that the entity definitions are available in the parametric memory of the LLM. In the first setting, we prompt the language model to decide whether or not a pair of entities have a parent-child relationship (without using the definitions). We refer to this approach with entity-without-definition.

Grounding: with definition

In the second setting, **Step 1** uses a prompt that includes a definition of each entity and again, it is the language model to decide whether or not the pair of entities have a parent-child relationship. This provides contextual grounding and we expect the additional context to enhance performance [12]. We also consider the case of internal grounding - where the LLM is prompted to generate its own definition for each entity. We refer to this approach with grounding-with-definition.

Definition generation

Given the broad parametric knowledge of the LLM at hand, we conduct a third evaluation where the definitions are generated by the LLM itself and the taxonomy construction subsequently proceeds with the grounding-with-definition approach.

Section 2: Evaluation

Dataset

We use the zoology taxonomy to evaluate the different approaches. The taxonomy structure and taxon definition data is publicly available from wikipedia[1] and university of michigan[4]. We need only a small subset of the full zoology taxonomy (see appendix II), to illustrate the benefits of the method presented.
We use the python networkx package to build the parent-child graph and prune it using transitive reduction[3]. We use the gemini flash 2.5 api to prompt the LLM[6] (see Appendix I for the detailed prompts).

Experimental results

We run the taxonomy reconstruction using the prompts shown in the Appendix. We vary the LLM temperature parameter over the values: 0.5, 1, 1.5 (the full range is from 0 to 2 [6]) to assess the variability of the parent-child linking decision depending on the temperature. Using the entity-name-only method, we find that prompt based parent-child extraction of the zoology taxonomy, which is included in the training data of the model, shows encouraging accuracy but poor reliability (Figure 2). When the correct parent-child relationships are identified we get an acyclic graph and the transitive reduction results in the correct taxonomy. The count of acyclic graphs is constant when using definitions but shows significant variability otherwise (Figure 1).

Figure 1

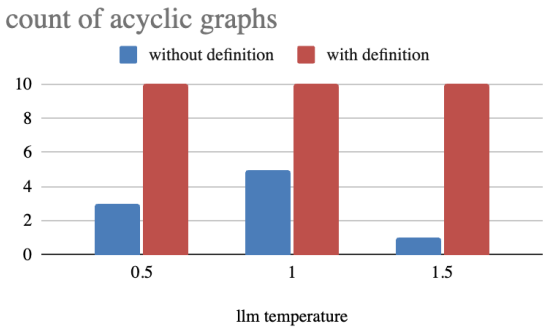


Figure 2

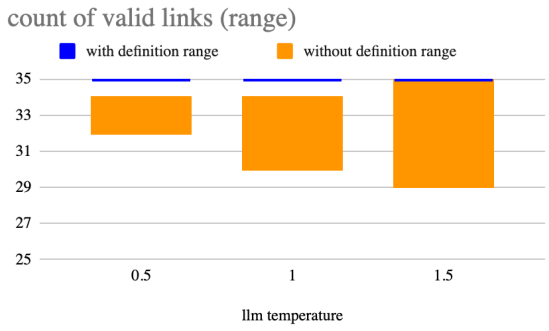


Figure 3

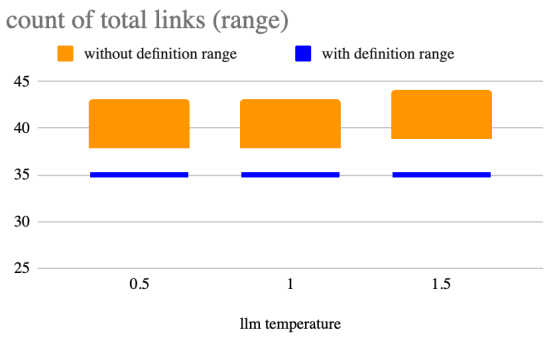
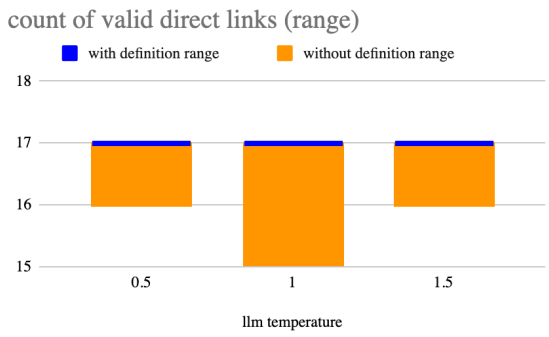


Figure 4



Using the grounding-with-definition, we observe that the parent-child relationships are more accurate and much more stable (Figure 3). In this small example, all of the 17 direct parent-child relationships are recovered and all indirect (e.g. grand-parent-child) relationships are correct (Figure 4). In Step 2, the transitive reduction consistently results in the correct taxonomy.

We conduct a third and final experiment where the LLM is prompted to generate the definition for each taxon. The generated definitions are cached (each taxon's definition is generated only once per experiment) and used in the grounding-with-definition process. **The results are identical to the initial grounding-with-definition experiment**. This clearly

demonstrates that the LLM, while having the parametric knowledge, may fail to use it. **It also confirms that bringing entity definitions as part of the prompt, i.e. in context grounding, while constructing the taxonomy improves results' stability and accuracy.**

Section 3: Related work

Parent-child or hypernym relationship detection was framed as a task on pairs of words with well established datasets such as BLESS [7]. To improve performance on the BLESS dataset and others, syntactic patterns for hypernyms are exploited in conjunction with NLP parsers and LLMs [8][9]. This research relies solely on the LLMs inherent capability to analyse text and identify parent-child relationships. Camacho-Collados [10] defines the task of identifying hypernyms from a corpus, given a list of names. While we limit ourselves to enriching our working context only with the definition of the names in the potential parent-child pair. Here we leverage only contextual grounding while [13] proposes grounding through model adaptation. Regardless of the strength of grounding, definitions that are ambiguous or self-contradicting will still result in unreliable results, analogous to the case entity-without-definition. The rating of parent-child relationship decisions is left for future work. Detecting uncertain LLM decisions for example using [14] could provide the opportunity to bring a human in the loop to interactively improve the problematic definitions. Improving **Step 2**, the extraction of the taxonomy tree out of the directed graph resulting from the parent-child links is out of the scope of this paper.

As future research, we could extend the scope of this experiment using the full zoology taxonomy and other domain specific taxonomies to evaluate the scalability and generality of this approach.

Conclusion

We presented an effective approach for taxonomy construction given taxa definition, entities definitions or glossary. We considered the case where the entities are part of the LLM training data. First we recognized that pre-trained models may fail in recognizing relationships that are part of their (vast) training data. Then we demonstrated this issue can be largely overcome by including the definition of the relevant entities in the prompt as in context learning while identifying parent-child relationships with an LLM. Finally we used the transitive reduction method to effectively prune redundant parent-child relationships of the obtained taxonomy. While this method remains sensitive to the quality of the definitions and the (continuously improving) grounding ability of the LLM[12], the combination of in context learning of definition and hypernymy relationship pruning with transitive reduction provides an effective method for automatic taxonomy construction.

References

[1] https://en.wikipedia.org/wiki/Taxonomy_(biology)

[2] https://en.wikipedia.org/wiki/International_Code_of_Zoological_Nomenclature

[3] https://en.wikipedia.org/wiki/Transitive_reduction

[4] https://animaldiversity.org/

[5] https://networkx.org/

[6] https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash-lite

[7] How we BLESSed distributional semantic evaluation (Marco Baroni, Alessandro Lenci, GEMS 2011)

[8] Exploring Prompt-Based Methods for Zero-Shot Hypernym Prediction with Large Language Models (Mikhail Tikhomirov, Natalia Loukachevitch, arXiv:2401.04515v1)

[9] Learning syntactic patterns for automatic hypernym discovery (Snow, Rion, Daniel Jurafsky, and Andrew Y. Ng., NIPS 2004)

[10] SemEval-2018 Task 9: Hypernym Discovery (Camacho-Collados et al., SemEval 2018)

[11] A Critical Analysis of the Largest Source for Generative AI Training Data: Common Crawl (Stefan Baack, FAccT 2024)

[12] How Well Do Large Language Models Truly Ground? (Hyunji Lee, Sejune Joo, Chaeeun Kim, Joel Jang, Doyoung Kim, Kyoung-Woon On, Minjoon Seo, NAACL2022)

[13] Effective Large Language Model Adaptation for Improved Grounding and Citation Generation (Xi Ye, Ruoxi Sun, Sercan Ö. Arik, Tomas Pfister, NAACL2024)

[14] Kernel Language Entropy: Fine-grained Uncertainty Quantification for LLMs from Semantic Similarities (Alexander Nikitin, Jannik Kossen, Yarin Gal, Pekka Marttinen, arXiv:2405.20003v1, 2024)

Appendix I: Prompt templates

```
# prompt templates

# definition generation
generate_definition_prompt = "Give a textual definition of {} with 250 words or less." + '\n'

generate_defintion_response_format = " Respond in json format for example: ```json
{'definition': 'Chordata is a primary phylum within the kingdom Animalia.'}```"

# definition prompt
definition_prompt = "Definition of '{}': {}" + '\n'

# parent-child relationship identification prompt
is_type_of_relationship = "Has '{}' all parts and attributes and is a specific instance of (is-a
relationship, an hyponym of) '{}' ?" + '\n'

response_prompt = " Respond in json format for example: ```json {'is_a': 'true'}```"

 right_and_left_definition_prompt = definition_prompt.format(right, right_definition)  +
definition_prompt.format(left, left_definition)

 is_right_a_type_of_left_prompt = right_and_left_definition_prompt +
is_type_of_relationship.format(right, left) + response_prompt
```

The complete code, including reading the dataset, gemini api call, json parsing, networkx graph building and reduction is available on request to the corresponding author.

Appendix II: Zoology taxonomy subset

Table 1: taxonomy direct and indirect links correctly identified by grounding-with-definition

child node	parent node	name pair	valid link	direct link
0	4	Amphibia Chordata	1	0
0	16	Amphibia Vertebrata	1	1
0	17	Amphibia Animalia	1	0
1	5	Anthozoa Cnidaria	1	1
1	17	Anthozoa Animalia	1	0
2	4	Aves Chordata	1	0
2	16	Aves Vertebrata	1	1
2	17	Aves Animalia	1	0
3	11	Cephalopoda Mollusca	1	1
3	17	Cephalopoda Animalia	1	0
4	17	Chordata Animalia	1	1
5	17	Cnidaria Animalia	1	1
6	17	Ctenophora Animalia	1	1
7	12	Demospongiae Porifera	1	1
7	17	Demospongiae Animalia	1	0
8	12	Hexactinellida Porifera	1	1
8	17	Hexactinellida Animalia	1	0
9	5	Hydrozoa Cnidaria	1	1
9	17	Hydrozoa Animalia	1	0
10	4	Mammalia Chordata	1	0
10	16	Mammalia Vertebrata	1	1
10	17	Mammalia Animalia	1	0
11	17	Mollusca Animalia	1	1
12	17	Porifera Animalia	1	1
13	4	Reptilia Chordata	1	0
13	16	Reptilia Vertebrata	1	1
13	17	Reptilia Animalia	1	0
14	6	Tentaculata Ctenophora	1	1
14	17	Tentaculata Animalia	1	0
15	4	Testudines Chordata	1	0
15	13	Testudines Reptilia	1	1
15	16	Testudines Vertebrata	1	0
15	17	Testudines Animalia	1	0
16	4	Vertebrata Chordata	1	1
16	17	Vertebrata Animalia	1	0
TOTAL			35	17

All the direct links are correctly identified. While each of the indirect link identified is also correct.