Development of in-context methods with Large Language Models for statement validation: towards semi-automated ontology learning in novel biological contexts

James Wilsenach[1,2,*], Sebastian Ahnert[1,3]

[1]The Alan Turing Institute, British Library, 96 Euston Road, London, NW1 2DB, United Kingdom
[2]The Earlham Institute,Norwich Research Park, Colney Lane, Norwich, NR4 7UZ, United Kingdom
[3]Department of Chemical Engineering and Biotechnology, University of Cambridge, Philippa Fawcett Drive, Cambridge, CB3 0AS, United Kingdom

Abstract

Biological knowledge is often contextual, requiring domain-specific expertise in order to determine the relative validity of a given statement, resulting in ever more context-specific ontologies. At the level of individual annotated entities the completeness of the annotational space is also contextual. For example, differences in both the number and depth of GO annotations are evident between model and non-model organisms. In the curation of domain-specific ontologies, parts of other hierarchies are often borrowed from a variety of related ontological sources. Learning which ontological statements can be imported from one context to another is therefore relevant to the goal of completing the ontological and annotational spaces and performing more accurate downstream analyses in a given context. We explore the use of inferred probabilities from light-weight GPT2-type Large Language Models to determine the contextual relevance of statements. We show that context-specific prompting can improve the ability of non-fine-tuned models to determine the correct direction in subtype relations for a simple taxonomy drawn from FOODON, the edible food ontology. This ontology was tested because of it's mixture of in-sample and out-of sample terms. The resulting measure could be used to gauge which statements are most contextually relevant when a model is appropriately prompted or fine-tuned. Through refining our neurosymbolic approach, we plan to provide a tool to guide investigators when deciding which ontological statements might plausibly be imported, such as from one species or cell line to another, and to identify possible areas for further study, where statements are least likely to hold.

Keywords

machine learning, large language models, statement validation, ontology validation, ontology comparison

1. Introduction: The LLM Inference Problem

The use of LLMs in the generation of ontological statements has been explored in a number of domains [1, 2]. A general form for the inference problem for ontological assertions is to consider the probability

$$\Pr(x \overset{R}{\sim} y|D) = \Pr(x|D)\Pr(\overset{R}{\sim}|x, D)\Pr(y|x \overset{R}{\sim}, D) \tag{1}$$

where x and y are ontological entities and $\overset{R}{\sim}$ represents a relation. Probabilities are also dependent on the specific domain D.

Figure 1 shows two sketches of ontologies for related domains, one describing the subclass relations in a hypothetical English *folk* ontology of plant-produce which separates fruits and vegetables somewhat arbitrarily for sociocultural reasons. In the second sketch, entities are separated according to the standards of FOODON, the edible food ontology [3], with all plant-produce including fruits a subclass of vegetable. This example illustrates the importance of domain specificity in using LLMs for the generation and recall of ontological knowledge even when ontologies are describing relations between similar entities.

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
*Corresponding author.
jwilsenach@turing.ac.uk (J. Wilsenach)
0000-0001-8214-9009 (J. Wilsenach); 0000-0003-2613-0041 (S. Ahnert)

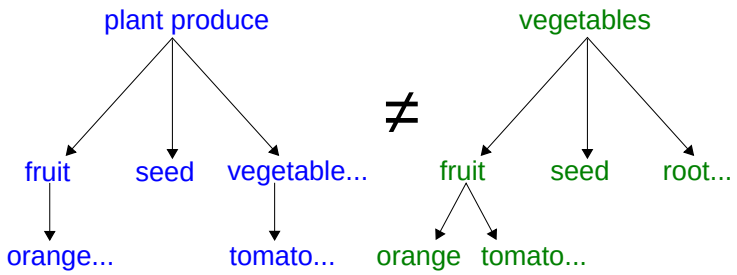

Figure 1: Examples of two ontologies based on a folk understanding of plant produce (left) and a formal and harmonized understanding posited in FOODON.

2. Context Improves Recall of Ontological Statements

We used a reduced taxonomy, termed FRUITON, consisting of all descendents of *fruit* in the FOODON ontology with 660 entities and 781 basic facts. We tested the recall of the light-weight version of DistilGPT2 (with 82M parameters) [4] on either true or false assertions in FRUITON by comparing the two posterior probabilities

$$\Pr_L(c \overset{\text{is\_a}}{\sim} p|C) \qquad \text{and} \qquad \Pr_L(p \overset{\text{is\_a}}{\sim} c|C)$$

which describes the probability of observing the basic fact $c \overset{\text{is\_a}}{\sim} p$ about the child entity c and parent p as well as the false assertion $c \overset{\text{is\_a}}{\sim} p$. We examine both assertions in the context of raw text C which appears before the statement of the assertion in the text and is either empty (context free) or provides additional domain-specific text, standing in for the domain D in Equation 1. This probability also depends on the language model L.

Our results show an increase in performance when providing domain-specific context with an AUC of 0.88 versus 0.85 for context free inference on the recall task. Figure 2 shows the results of LLM inference for a specific subset of FRUITON centred on *orange* when using domain-specific context. The model correctly identifies most correct and incorrect assertions.

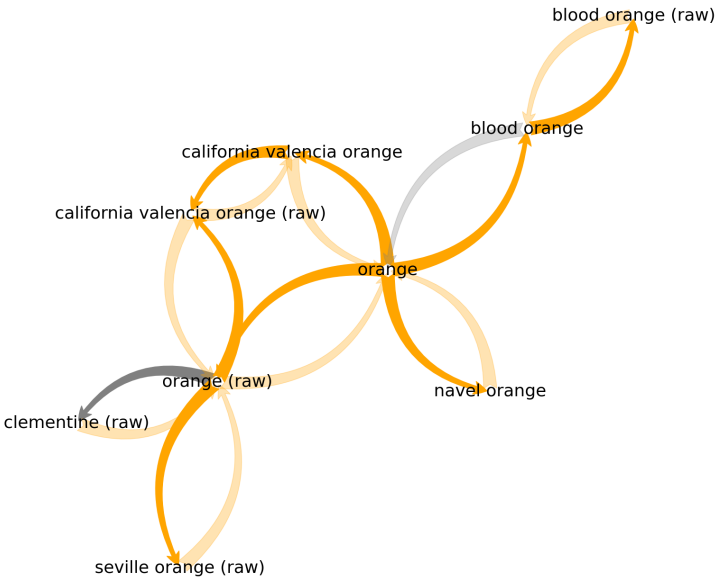

Figure 2: This figure shows a subgraph of FRUITON determined by the term *orange* and its descendents. Shaded edges between terms are shown in both the true parent to child direction (solid colour) and false child to parent, directions (transparent). Incorrectly identified edges at the given threshold of $t = 10^{-6}$ are shown in grey and correctly identified edges are shown in orange.

3. Future Directions: Neurosymbolic Approaches

The above recall task suggests a natural extension

$$\Pr_L(c \overset{R}{\sim} p^*|C) \qquad \text{and} \qquad \Pr_L(c^* \overset{R}{\sim} p|C)$$

where p^* and c^* represent entities that appear in valid assertions about c and p, including those that are entailed and can be deduced from computational reasoning. This provides both a broader corpus to test validity and a framework for recall testing such as proposed in [5]. This framework represents an exciting new direction for using scalable domain-specific biomedical LLMs in combination with novel measures to test the validity of assertions in one ontology in the context of novel domains.

Declaration on Generative AI

The authors have not employed any Generative AI tools outside of model design and evaluation.

References

[1] H. T. Mai, C. X. Chu, H. Paulheim, Do LLMs really adapt to domains? an ontology learning perspective, in: International Semantic Web Conference, Springer, 2024, pp. 126–143.

[2] S. Wadhwa, S. Amir, B. C. Wallace, Revisiting relation extraction in the era of large language models, in: Association for Computational Linguistics Proceedings, volume 2023, NIH Public Access, 2023, p. 15566.

[3] D. M. Dooley, E. J. Griffiths, G. S. Gosal, P. L. Buttigieg, R. Hoehndorf, M. C. Lange, L. M. Schriml, F. S. Brinkman, W. W. Hsiao, FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration, npj Science of Food 2 (2018) 23.

[4] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, in: NeurIPS EMC2 Workshop, 2019.

[5] T.-D. Bradley, J. Terilla, Y. Vlassopoulos, An enriched category theory of language: from syntax to semantics, La Matematica 1 (2022) 551–580.