

Beyond Similarity: LLM-Based Generation of Typed Semantic Links for Ontology Networks

Nouha Hayouni¹, Sheeba Samuel^{2,*} and Alsayed Algergawy³

¹Chair of Data and Knowledge Engineering, University of Passau, Germany

²Distributed and Self-organizing Systems, Chemnitz University of Technology, Germany

³FLZ Karlsruhe – Leibniz Institute for Information Infrastructure, Eggenstein-Leopoldshafen, and Karlsruhe Institute of Technology (AIFB), Germany

Abstract

Ontology networks enable interoperability by connecting independently developed ontologies through meaningful semantic relationships, but curating such links manually does not scale to large interdisciplinary collections. This paper presents a lightweight framework that assists ontology network construction by combining contextual language representations with large language model (LLM) reasoning. The workflow first narrows the search space through domain-adapted DistilBERT embeddings and clustering-based candidate selection before employing GPT-4o to infer typed semantic relationships together with textual justifications, which are then mapped onto operational semantic roles. Applied to ReproduceMeON, a 33-ontology network spanning computational science, machine learning, microscopy, and experimental workflows, the filtering strategy reduces roughly 800,000 candidate concept pairs to about 95,000 pairs suitable for LLM analysis. Dual expert validation of 429 generated relationships yields an overall precision of 80.19% (91.49% on high-certainty annotations) with substantial inter-annotator agreement (Cohen's $\kappa = 0.723$). Against five similarity-based baselines, the framework improves F1 by 30.9 percentage points over the strongest baseline, while an ablation restricted to the pre-filtered, high-similarity candidates shows every similarity method scoring at chance level ($AUC \approx 0.5$) – evidence that contextual LLM reasoning, not embedding similarity, drives the precision gain.

Keywords

Knowledge Interoperability, Ontology Network, Large Language Model, Reproducibility, Ontology Alignment

1. Introduction

Knowledge-intensive scientific applications increasingly combine information originating from multiple research disciplines. Experimental protocols, computational methods, machine learning techniques, and microscopy workflows are frequently represented by separate ontologies that have been developed independently for their respective communities. While these ontologies provide detailed conceptual models within individual domains, integrating them into coherent knowledge infrastructures remains a significant challenge. *Ontology networks* (ONs) address this issue by establishing semantic connections across existing ontologies instead of attempting to replace them with a single comprehensive model [1, 2]. The creation of ONs depends on identifying meaningful relationships between concepts originating from different ontologies. Besides traditional equivalence mappings, practical ontology integration requires richer semantic associations describing dependencies, functionality, composition, causality, or workflow interactions. Constructing these links manually demands extensive domain expertise and quickly becomes impractical when thousands of ontology classes are involved.

Classical ontology matching systems [3] primarily focus on discovering equivalent entities, whereas embedding-based techniques such as OWL2Vec* [4] and RDF2Vec [5] estimate semantic proximity between concepts. Although these methods provide valuable support for candidate retrieval, similarity scores alone cannot reliably determine whether two concepts should be connected through a meaningful semantic relation – a claim our ablation study confirms empirically. Recent advances in LLMs have

SEMANTiCS 2026: International Conference on Semantic Systems, Posters and Demos Track (Fast-Track), 2026

*Corresponding author.

✉ nouha.hayouni@uni-passau.de (N. Hayouni); sheeba.samuel@informatik.tu-chemnitz.de (S. Samuel); alsayed.algergawy@fiz-karlsruhe.de (A. Algergawy)

ORCID 0000-0002-7981-8504 (S. Samuel); 0000-0002-8550-4720 (A. Algergawy)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

demonstrated strong capabilities for semantic reasoning and contextual interpretation, and prior work has begun exploring LLM-driven ontology and knowledge-graph construction [6], though typically without systematic candidate pre-filtering or role-based materialization of the resulting links.

Viewed abstractly, the decision problem we address resembles *link prediction* over class-level ontology concepts: given a candidate pair, decide whether a semantically meaningful relation holds and, if so, of which type. It differs from standard knowledge-graph link prediction in that the relation vocabulary is open and LLM-generated rather than fixed and learned from an existing graph, and it differs from classical alignment in targeting typed, non-equivalence relations rather than equivalence mappings – which is why neither family of methods forms a directly comparable baseline here.

Rather than relying solely on semantic similarity, the proposed framework employs a multi-stage workflow that first identifies promising candidate pairs using a domain-adapted DistilBERT [7] model and clustering-based filtering, and subsequently exploits GPT-4o to infer typed semantic relationships accompanied by human-readable justifications, which are further transformed into semantic roles for incorporation into ontology-driven scientific workflows. Code, data, and the expert-annotated validation set are available at https://github.com/fusion-jena/ReproduceMeON/tree/main/LLM_based_link_discovery.

2. Proposed Framework

The objective of the proposed framework is to automatically identify meaningful semantic relationships between concepts originating from multiple ontologies and to use these relationships for enriching ontology networks. Formally, given ontologies with class sets $C_1, \dots, C_n$, the goal is to discover typed links (c_a, r, c_b, j) with $c_a \in C_i$, $c_b \in C_j$, relation r drawn from a fixed inventory of non-equivalence relation types, and j a natural-language justification; a link is considered valid once a domain expert confirms that r holds between c_a and c_b . Equivalence mappings, handled by classical alignment tools, are explicitly out of scope. Instead of evaluating every possible concept pair, the framework progressively reduces the search space before applying semantic reasoning with a large language model. The framework consists of five sequential stages: ontology preprocessing, contextual representation learning, candidate selection, relationship generation, and semantic role assignment.

Ontology Preprocessing: The process begins by extracting semantic information from the ontologies contained in ReproduceMeON [2, 8]. Each ontology is parsed to obtain class-level RDF triples together with descriptive metadata such as labels, comments, and annotations. Since literals and instance-level assertions are not required for relationship discovery, only class-oriented information is retained. Incorporating textual descriptions at this stage provides additional contextual information that helps distinguish concepts with similar names but different meanings across scientific domains.

Contextual Representation Learning: To capture domain-specific semantics, ontology concepts are transformed into textual descriptions and used to fine-tune DistilBERT [7] with a masked language modeling objective, yielding dense contextual embeddings, richer than generic language embeddings, that encode both structural ontology information and descriptive annotations, and are more suitable for identifying semantically related entities across heterogeneous ontologies.

Candidate Selection: Performing relationship generation on every possible concept pair would be computationally expensive because the ontology network contains hundreds of thousands of potential combinations. Therefore, embeddings are first grouped into semantically coherent clusters using KMeans clustering, with the number of clusters chosen via the Elbow and Silhouette methods. Pairwise cosine similarity is computed only within each cluster, substantially reducing unnecessary comparisons. Candidate pairs are retained using a dynamic similarity threshold set at the 90th percentile of similarity values within each cluster. This strategy adapts to each cluster’s density, preserving highly relevant candidates while eliminating a large proportion of unrelated concept pairs before the LLM stage – in practice reducing the roughly 800k raw candidate pairs to about 95k (an 88% reduction), which is what makes the subsequent LLM step computationally feasible.

LLM-Based Relationship Generation: The filtered concept pairs are submitted to GPT-4o for

semantic interpretation. Besides the concept names, the prompt includes ontology provenance, domain information, and similarity scores, enabling the model to reason beyond lexical similarity. For each pair, the model determines whether a meaningful relationship exists and, when appropriate, assigns one of nine predefined semantic categories inspired by prior work on foundational and upper-level ontology relations [9]: hierarchical, functional, causal, temporal, compositional, instrumental, comparative, transformational, and spatial. In addition, GPT-4o generates a concise textual explanation describing why the relationship is appropriate, providing transparency for subsequent expert review.

To improve consistency, the prompts were iteratively refined through prompt engineering techniques such as explicit task instructions, structured output constraints, contextual cues, chain-of-thought decomposition [10], and few-shot examples. These refinements reduced ambiguous outputs (e.g., generic “is related to” relations) and encouraged more precise relationship descriptions.

Semantic Role Assignment: The final stage transforms the generated semantic relationships into operational roles that can be directly incorporated into the ontology network, grounded in established Ontology Design Patterns [11]. Depending on the inferred relationship, concepts are assigned roles such as *Provider*, *Consumer*, *Processor*, *Enabler*, or *Requester* – for instance, “DataSet Provides Data for Algorithm” instantiates a Provenance-style pattern and becomes directly queryable via SPARQL once materialized in the network.

3. Experimental Evaluation

The proposed framework was evaluated using ReproduceMeON [2, 8], an ontology network comprising 33 ontologies from the domains of computational science, machine learning, microscopy, and experimental workflows, to test whether combining contextual embeddings with LLM reasoning discovers meaningful semantic relationships that similarity measures alone cannot identify.

Experimental Design: Experiments were run using the Hugging Face Transformers, PyTorch, RDFlib, and Scikit-learn libraries, with GPT-4o accessed via the OpenAI API for relationship generation.

Relationship Generation: The framework generated both intra-domain and inter-domain semantic links spanning multiple scientific disciplines, with a near-even split (51.3% inter-domain, 48.7% intra-domain). Functional relationships constituted the largest category, reflecting the workflow-oriented nature of reproducibility ontologies, while hierarchical, compositional, causal, comparative, transformational, and instrumental relations were also identified.

Comparison with Similarity-Based Methods: To assess the effectiveness of the proposed approach, we compared it against five similarity-based techniques: Jaccard similarity on concept names, Jaccard similarity with contextual metadata, TF-IDF cosine similarity, pre-trained (non-fine-tuned) DistilBERT embeddings, and Sentence-BERT. This tests whether semantic similarity alone suffices to judge relationship validity; it is not a claim of outperforming equivalence-mapping systems such as AML or LogMap, which address a different task. The evaluation set comprises the 344 valid and 85 invalid expert-annotated relationships plus 500 randomly generated cross-domain pairs as easy negatives; since random negatives are easier to reject than real hard negatives, the resulting F1 scores should be read as performance on this constructed benchmark rather than a real-world matching task – expert-verified **Precision = 80.19%** on the actual pipeline output remains the primary metric.

Table 1 reports Precision/Recall/F1/AUC (Area Under the ROC Curve) on the full evaluation set, alongside AUC restricted to the 429 pipeline pairs already filtered by the 90th-percentile similarity threshold – if similarity alone were sufficient, this restricted AUC should approach 1.0.

The proposed framework outperformed every baseline, with F1 rising 30.9 percentage points over the strongest one (Sentence-BERT). More strikingly, every similarity method scores at chance level ($AUC_{429} \approx 0.5$) on the pre-filtered pairs, regardless of whether it uses surface-level lexical features (Jaccard, TF-IDF) or deep contextual embeddings (DistilBERT, Sentence-BERT): among candidates that already share high similarity, similarity carries no further signal about validity, and separating valid from invalid links requires the contextual reasoning the LLM step contributes.

Table 1

Performance comparison with similarity-based baselines. AUC_{929} : full 929-pair evaluation set. AUC_{429} : restricted to the 429 pre-filtered, high-similarity pipeline pairs.

Method	Precision	Recall	F1	AUC_{929}	AUC_{429}
Jaccard (names)	0.370	1.000	0.540	0.506	0.506
Jaccard (context)	0.370	1.000	0.540	0.574	0.485
TF-IDF	0.576	0.552	0.564	0.625	0.503
DistilBERT	0.582	0.515	0.546	0.609	0.511
Sentence-BERT	0.570	0.593	0.581	0.673	0.500
Proposed Framework	0.802	1.000	0.890	0.927	–

3.1. Expert Validation:

To assess the quality of the generated semantic links, two domain-expert annotators manually evaluated a sample of 429 relationships produced by the framework. Annotator 1 covered all 429 relationships across all four domains; Annotator 2 independently covered 274 of them across three domains, so the Microscopy domain (155 pairs) has single-annotator coverage only. Overall, 344 relationships were judged to be correct, corresponding to an expert precision of 80.19%; restricting to high-certainty annotations raises this to 91.49% – expected, since high-certainty pairs are by definition the less ambiguous cases. On the 274 dually annotated pairs, agreement between the two annotators reaches Cohen’s $\kappa = 0.723$ (90.5% raw agreement) – *substantial agreement* – ranging from $\kappa = 0.671$ (Machine Learning) to $\kappa = 0.817$ (Computational) across domains. Table 2 breaks the expert-validated precision down by domain and link type.

Table 2

Precision of generated relationships by domain and link type.

Domain	Valid	Total	Precision
Machine Learning	103	137	75.18%
Microscopy	126	155	81.29%
Computational	80	94	85.11%
Experimental Workflow	35	43	81.40%
Overall	344	429	80.19%
Intra-domain	192	227	84.58%
Inter-domain	152	202	75.25%

Intra-domain precision (84.58%) exceeds inter-domain precision (75.25%), consistent with intra-domain pairs sharing more vocabulary and being easier to align semantically. By relation type, *Causal* and *Comparative* relations reach 100% precision and *Hierarchical* reaches 95.24%, while *Instrumental* relations are the most error-prone (65.22%); the dominant failure mode is inter-domain pairs where lexical overlap masks conceptual incompatibility (e.g., machine-learning process terms linked to microscopy procedural terms), with the model often conflating a tool-role entity with a process-role entity.

Limitations. Recall cannot be computed, since no exhaustive gold standard of all possible cross-ontology links exists; the balanced evaluation set uses easy random negatives, which inflates F1 relative to real hard negatives (Precision on pipeline output is unaffected by this); the second annotator covered 274 of 429 pairs, with Microscopy annotated by a single expert; all experiments use one ontology network and one LLM (GPT-4o), whose stochastic outputs mean repeated runs may yield different labels; and the semantic-role assignment’s correctness has not been independently evaluated, e.g., via competency questions.

4. Conclusion

This paper presented a framework for supporting ontology network construction through the automatic generation of typed semantic relationships between ontology concepts, combining domain-adapted DistilBERT embeddings, clustering-based candidate selection, GPT-4o-driven relationship generation, and ODP-grounded role mapping. Applied to ReproduceMeON’s four domains, dual expert validation of 429 relationships yields 80.19% precision (91.49% on high-certainty annotations) with substantial inter-annotator agreement ($\kappa = 0.723$). Against five similarity-based baselines, the framework reaches F1 = 0.890 versus the best baseline’s 0.581, and an ablation restricted to the 429 pre-filtered, high-similarity pairs shows every similarity method at chance-level AUC, confirming that similarity alone cannot judge relationship validity and that LLM-based reasoning over concept roles and domain semantics is what delivers the precision gain. Embedding-based pre-filtering remains essential for feasibility, cutting roughly 800k candidate pairs to about 95k. The proposed workflow is intentionally modular and can be adapted to other ontology collections or alternative language models. Future work includes comparing open-source instruction-tuned LLMs, curating a gold-standard subset to measure recall, extending dual annotation to the Microscopy domain, applying the framework to a second ontology network to test generalizability, and evaluating the semantic role mapping through competency questions and SPARQL-based querying combined with OWL reasoning over the enriched network.

Declaration on Generative AI During the preparation of this work, the authors used ChatGPT and Claude to improve the text of the manuscript. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

- [1] P. Haase, S. Rudolph, Y. Wang, S. Brockmans, D1. 1.1 networked ontology model, 2006.
- [2] S. Samuel, B. König-Ries, A. Algergawy, The role of ontology matching in ontology network development, in: Proceedings of the 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference, 2023. URL: https://ceur-ws.org/Vol-3591/om2023_LTpaper2.pdf.
- [3] J. Euzenat, P. Shvaiko, Ontology Matching: State of the Art and Future Challenges, 2nd ed., Springer, 2013. URL: <https://doi.org/10.1109/TKDE.2011.253>.
- [4] J. Chen, P. Hu, E. Jimenez-Ruiz, O. M. Holter, D. Antonyrajah, I. Horrocks, Owl2vec*: Embedding of owl ontologies, Machine Learning 110 (2021) 1813–1845.
- [5] P. Ristoski, H. Paulheim, Rdf2vec: Rdf graph embeddings for data mining, in: The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part I 15, Springer, 2016, pp. 498–514.
- [6] V. K. Kommineni, B. König-Ries, S. Samuel, From human experts to machines: An LLM supported approach to ontology and knowledge graph construction, CoRR abs/2403.08345 (2024). URL: <https://doi.org/10.48550/arXiv.2403.08345>. doi:10.48550/ARXIV.2403.08345.
- [7] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019). URL: <https://arxiv.org/abs/1910.01108>.
- [8] S. Samuel, A. Algergawy, B. König-Ries, Towards an ontology network for the reproducibility of scientific studies 8 (2021) 1–12. URL: <https://ceur-ws.org/Vol-2969/paper64-OntoCom.pdf>.
- [9] I. Niles, A. Pease, Towards a standard upper ontology, FOIS ’01, Association for Computing Machinery, New York, NY, USA, 2001, p. 2–9. doi:10.1145/505168.505170.
- [10] J. Wei, et al., Chain of thought prompting elicits reasoning in large language models, in: Proceedings of the Neural Information Processing Systems, 2022. URL: <https://arxiv.org/abs/2201.11903>.
- [11] A. Gangemi, Ontology design patterns for semantic web content, in: Proceedings of the 4th International Semantic Web Conference (ISWC 2005), volume 3729 of *Lecture Notes in Computer Science*, Springer, 2005, pp. 262–276.