Ontology Text Alignment: Aligning Textual Content to Terminological Axioms (Abstract)*

Jieying Chen[1,2], Hang Dong[1,3], Jiaoyan Chen[1,4], and Ian Horrocks[1]

[1] University of Oxford, UK
firstname.lastname@cs.ox.ac.uk
[2] Vrije Universiteit Amsterdam, Netherlands
j.y.chen@vu.nl
[3] University of Exeter, UK
H.Dong2@exeter.ac.uk
[4] University of Manchester, UK
jiaoyan.chen@manchester.ac.uk

Despite recent advancements in Large Language Models (LLMs), challenges persist in their ability to perform reasoning and provide explainable outcomes, highlighting the ongoing importance of ontologies in certain domains. However, ontology modelling remains complex, requiring extensive human expertise and effort. To address these ongoing challenges in ontology engineering, methods such as ontology modularity and ontology alignment have been developed to simplify the creation of more functional subsets of ontologies. However, these techniques often present challenges, particularly because it is difficult for users to accurately identify the signature, the set of concept and role names that users are interested in. An imprecise signature can lead to erroneous results. Furthermore, controlling the size of the modules [1–3,5–8,11,14] directly is not feasible, adding another layer of complexity to the process. Additionally, in real-world applications, ontologies frequently need to be crafted from academic research and other sources. This reality led us to introduce the problem of *Ontology Text Alignment*. Our proposed solution to this problem addresses this problem and has broad applications, including enhancing ontology modelling, boosting semantic search capabilities, and refining the ontology selection process.

Unlike traditional named entity recognition and entity linking problems, our problem focuses on inferring the text's claims from axioms rather than simply identifying concept mentions. This adds complexity and increases the challenge. Formally, the definition of the problem is defined as follows:

Definition 1 (Ontology Text Alignment). *Let $\mathcal{O}$ be an ontology, $\mathcal{E} \subseteq \mathcal{O}$, ref be the reference text, and let $k \in \mathbb{Z}^+$. Additionally, let $\mu : (\mathcal{O}, \mathcal{E}, \mathsf{ref}) \mapsto \mathbb{R}_{\geq 0}$ be a relevance measure function. The task* ontology text alignment *w.r.t. ref under μ is to identify an $\mathcal{E} \subseteq \mathcal{O}$ satisfying the following condition:*

$$\mu(\mathcal{O}, \mathcal{E}, \mathsf{ref}) = \max\{\, \mu(\mathcal{O}, \mathcal{E}', \mathsf{ref}) \mid \mathcal{E}' \subseteq \mathcal{O}, |\mathcal{E}'| \leq k \,\}.$$

The objective is to identify a subontology that achieves the highest relevance score w.r.t. the given reference text, subject to the size constraint k. Note that

* The full paper was accepted by ECAI 2024.

our paper primarily focuses on terminological axioms (TBox axioms) characterized by their rich semantics and complex structures. We do not consider factual assertions, namely, class/role assertions (ABox axioms).

Framework. We tackle this problem by identifying the most relevant, top-ranked axioms, inspired by recent advancements in generative LLMs and the Retrieval-Augmented Generation framework. Our approach includes the following components, based on these key considerations:

- **Verbalization.** Complex axioms often contain complex logical operations like conjunctions and existential restrictions, and classes are encoded as URLs, making them challenging and not inherently readable by current LLMs. Through verbalization, we transform axioms into descriptive natural language texts.
- **Indexing via Axiom Text Embedding.** Handling large ontologies is challenging for current generative LLMs due to their limitations with long texts. We tackle this by utilizing pre-trained models such as BERT [4] for initial indexing. Specifically, we use BERT models to convert axiom sentences and reference texts into vectors, and then apply cosine similarity to identify and rank the most relevant texts.
- **Semantic enrichment.** In the last stage, the internal structure and semantics of the ontology were not considerred. To address this, we propose enhancing the top-ranked axioms by incorporating the ontology's internal structure. We achieve this by constructing ontology atomic graphs using atom decomposition theory [10, 15] to refine the retrieval results.
- **Integration of generative LLMs.** We aim to improve the accuracy and relevance of indexing by integrating semantically rich ontology graphs into LLM prompts. This enriches the indexing process with contextually relevant knowledge, enhancing the understanding and use of ontological structures via atomic decomposition.

Benchmark Creation we developed three benchmark datasets across diverse domains: geology, food, and medicine. These datasets are based on extractive summaries derived from three specific ontologies: GeoFault [12], and two branches of the SNOMED CT ontology that focus on diseases and anatomy.

Evaluation We employ several BERT-based models for token embeddings, including the basic BERT model, SBERT [13], and SapBERT [9]. For generative LLMs, we utilize GPT 3.5 and 4 series, LLaMA 2 7b and 13b versions. Baseline comparisons include TF-IDF and Word2Vec embeddings, aggregated via mean pooling, with the similarity quantified through cosine similarity. Our evaluation shows that the integration of LLMs, like GPT models, together with semantic enrichment components significantly boosts our framework's performance, especially with configurations such as SBERT and SapBERT. For instance, upgrading from version 3.5 to 4 results in substantial improvements, highlighting the effectiveness of these advanced models in semantic tasks.

References

1. Chen, J., Ludwig, M., Ma, Y., Walther, D.: Towards extracting ontology excerpts. In: Proc. of KSEM'15: the 8th International Conference on Knowledge Science, Engineering and Management. pp. 78–89 (2015)
2. Chen, J., Ludwig, M., Ma, Y., Walther, D.: Computing minimal projection modules for *ELH*ˆr -terminologies. In: Calimeri, F., Leone, N., Manna, M. (eds.) Logics in Artificial Intelligence - 16th European Conference, JELIA 2019, Rende, Italy, May 7-11, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11468, pp. 355–370. Springer (2019). https://doi.org/10.1007/978-3-030-19570-0_23
3. Chen, J., Ludwig, M., Walther, D.: Computing minimal subsumption modules of ontologies. In: Proc. of GCAI'18. pp. 41–53 (2018)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423
5. Grau, B.C., Horrocks, I., Kazakov, Y., Sattler, U.: Modular reuse of ontologies: Theory and practice. J. Artif. Intell. Res. **31**(1), 273–318 (2008)
6. Konev, B., Lutz, C., Walther, D., Wolter, F.: Model-theoretic inseparability and modularity of description logic ontologies. Artif. Intell. **203**, 66–103 (2013)
7. Kontchakov, R., Wolter, F., Zakharyaschev, M.: Logic-based ontology comparison and module extraction, with an application to dl-lite. Artif. Intell. **174**(15), 1093–1141 (2010)
8. Koopmann, P., Chen, J.: Deductive module extraction for expressive description logics. In: Bessiere, C. (ed.) Proceedings of IJCAI'20. pp. 1636–1643. ijcai.org (2020)
9. Liu, F., Vlachos, A., Cohn, T.: Self-alignment pretraining for biomedical entity representations. Nature Machine Intelligence **3**(4), 316–325 (2021)
10. Martín-Recuerda, F., Walther, D.: Fast modularisation and atomic decomposition of ontologies using axiom dependency hypergraphs. In: Proc. of ISWC'14: the 13th International Semantic Web Conference. Lecture Notes in Computer Science, vol. 8797, pp. 49–64 (2014)
11. Mossakowski, T., Codescu, M., Neuhaus, F., Kutz, O.: The distributed ontology, modeling and specification language – dol. In: Koslow, A., Buchsbaum, A. (eds.) The Road to Universal Logic, vol. 2, pp. 489–520. Birkhäuser (2015)
12. Qu, Y., Perrin, M., Torabi, A., Abel, M., Giese, M.: Geofault: A well-founded fault ontology for interoperability in geological modeling. Comput. Geosci. **182**, 105478 (2024). https://doi.org/10.1016/J.CAGEO.2023.105478
13. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. pp. 3982–3992 (2019)
14. Stuckenschmidt, H., Parent, C., Spaccapietra, S.: Modular Ontologies: Concepts, Theories and Techniques for Knowledge Modularization, Lecture Notes in Computer Science, vol. 5445. Springer Verlag (01 2009)
15. Vescovo, C.D., Parsia, B., Sattler, U., Schneider, T.: The modular structure of an ontology: Atomic decomposition. In: Proc. of IJCAI'11: the 22nd International Joint Conference on Artificial Intelligence. pp. 2232–2237. IJCAI/AAAI (2011)