Confirmation of Status Report

For thesis tentatively titled: *Language Model for Ontology Engineering*

Yuan He

St Hugh's College

Department of Computer Science

University of Oxford

A progress report submitted for the degree of

Doctor of Philosophy

Trinity 2023

Supervisors

Prof. Ian Horrocks
Prof. Bernardo Cuenca Grau
Dr. Jiaoyan Chen

1 Introduction

Ontology engineering is a sub-field of knowledge engineering that involves different stages of ontology development (Staab and Studer, 2010), such as ontology design, construction, curation, evaluation, maintenance, and more. An ontology is a formal, explicit specification of knowledge within the scope of a domain. It provides a vocabulary of concepts and properties that enables a shared understanding of semantics among humans and machines. Tasks of ontology engineering include: *(i)* defining the entities and constructing the logical axioms that compose the ontology, *(ii)* validating and ensuring quality (e.g., completeness and correctness) of the built ontology, *(iii)* inserting new knowledge into the ontology, *(iv)* integrating domain ontologies that come from heterogeneous sources, and so forth. These activities collectively ensure that an ontology is applicable and relevant to real-world scenarios, enhancing its practical utility.

Language models (LMs), having emerged as a significant force in AI research, have been increasingly applied to ontology engineering. Techniques based on LMs exhibit noticeable advantages over traditional, symbolic, or rule-based ontology engineering tools. For example, LogMap (Jiménez-Ruiz and Cuenca Grau, 2011), an eminent ontology alignment system, relies on lexical similarity and falls short in capturing contextual text embeddings. In contrast, BERTMap (He et al., 2022a), an LM-based system, exploits the self attention mechanism of the transformer architecture (Vaswani et al., 2017), demonstrating robustness to linguistic variations such as synonyms and polysemies. Pre-trained on extensive text corpora, LMs come equipped with a rich background knowledge, which allows them to handle complex tasks such as ontology construction, requiring only carefully designed prompts to guide their responses (Caufield et al., 2023).

Despite their strengths, LMs are often challenged in providing clear and explicit reasoning behind their responses due to their inherent focus on learning distributional patterns from training texts. To capture logical or structural information, which is particularly important in ontology engineering, we can turn to models based on geometric embeddings. As an illustration, Abboud et al. (2020) proposed the usage of Euclidean hyper-rectangles (or Box Embeddings) for capturing the subsumption relationships among ontology concepts. Similarly, Lu et al. (2019) employed hyperbolic

Poincaré models to learn the hierarchy of medical terms derived from ontologies. Our research aims to integrate such geometric embeddings with LMs to devise a hybrid model of ontology embeddings. These embeddings will contain both the textual semantics commonly captured by LMs and the structural semantics inherent in geometric representations, and can hopefully be applied to a range of downstream ontology engineering tasks.

This progress report aims to provide an overview of the ongoing research and a roadmap for the conclusive D.Phil year. It will act as a guide for the forthcoming thesis provisionally titled: *Language Model for Ontology Engineering*.

2 Summary of Current Progress

As reflected by the tentative title, the central emphasis of the D.Phil research revolves around the exploration of how language models (LMs) can be harnessed for diverse tasks in ontology engineering. We are primarily confronted with two challenges:

(i) **Linearisation**, which is about transforming complex, structural data from ontologies into linear text sequences that LMs can handle.

(ii) **Adaptation**, which is to adjust or re-formulate task settings to match the training and inference paradigms of LMs.

As a starting point, we put forth BERTMap (He et al., 2022a), an LM-based ontology alignment system that takes two ontologies as inputs and yields class equivalence mappings between them. In this work, we began to tackle the second challenge by adopting a pipeline or a “divide-and-conquer” approach. This strategy subdivides the task into several smaller sub-tasks, employing the LM solely for synonym classification—a task at which the LM can achieve exceptionally high accuracy. The raw mappings are computed based on the ensemble results of the synonym classification probabilities, and then further refined through mapping extension and repair to form final output mappings.

Following BERTMap, we presented a related system, BERTSubs (Chen et al., 2022), which broadens the task to subsumption matching. In contrast to BERTMap, BERTSubs also addresses the first challenge by enhancing the representation of ontology classes with diverse graph-based[1] contexts. Moreover, we identified a disparity in the evaluation and comparison of symbolic and sub-symbolic ontology alignment systems. To rectify this, we introduced new datasets and a comprehensive evaluation framework in Bio-ML (He et al., 2022b).

However, both BERTMap and BERTSubs rely on substantial fine-tuning, which is typically less efficient and less generalisable. To delve deeper into both challenges, we developed OntoLAMA, aimed at probing an LM's comprehension of ontology semantics. OntoLAMA adopts the prompt learning paradigm, which formulates the subsumption inference task in a way akin to LM pre-training. This method aims to extract knowledge from an LM more effectively, without requiring excessive training. In addition, this study involves a recursive class verbaliser, which allows for testing the subsumption inference even for complex ontology classes.

The promising results derived from these studies demonstrate that, with suitable data linearisation and task adaptation, LMs can indeed be seamlessly integrated into an ontology engineering workflow. Furthermore, they establish a strong foundation for future exploration into more advanced topics, such as developing multi-purpose ontology embeddings and constructing ontologies directly from texts.

The works mentioned above are primary to the D.Phil research but do not cover all the projects completed thus far. A complete list of related publications, which provide further insights into the depth and breadth of the research carried out, can be found in the subsequent section.

3 Relevant Publications

Below is a list of relevant publications, organised chronologically. [1] — [3], and [5] are related to ontology alignment, [4] and [7] are about language model probing for knowledge, [6] and [7] both concern subsumption inference, and [8] is about entity linking and discovery.

[1] Here ”graph” refers to a structure with ontology classes as nodes and subsumption relations as edges.

[1] **Yuan He**, Jiaoyan Chen, Denvar Antonyrajah, and Ian Horrocks. “Biomedical Ontology Alignment with BERT”. The Sixteenth International Workshop on Ontology Matching (OM@ISWC'21) (He et al., 2021).

[2] **Yuan He**, Jiaoyan Chen, Denvar Antonyrajah, and Ian Horrocks. “BERTMap: A BERT-based Ontology Alignment System”. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI'22) (He et al., 2022a).

[3] **Yuan He**, Jiaoyan Chen, Hang Dong, Ernesto Jiménez-Ruiz, Ali Hadian and Ian Horrocks. “Machine Learning-Friendly Biomedical Datasets for Equivalence and Subsumption Ontology Matching”. In: The 21st International Semantic Web Conference (ISWC'22; **Best Resource Paper Candidate**) (He et al., 2022b).

[4] [2]Ryan Brate, Minh-Hoang Dang, Fabian Hoppe, **Yuan He**, Albert Meroño Peñuela, and Vijay Sadashivaiah. “Improving Language Model Predictions via Prompts Enriched with Knowledge Graphs”. In: Workshop on Deep Learning for Knowledge Graphs (DL4KG@ISWC'22) (Brate et al., 2022).

[5] [3]M. A. N. Pour, ..., **Yuan He**, ..., Lu Zhou. “Results of the Ontology Alignment Evaluation Initiative 2022”. The Seventeenth International Workshop on Ontology Matching (OM@ISWC'22) (Pour et al., 2022).

[6] Jiaoyan Chen, **Yuan He**, Ernesto Jiménez-Ruiz, Hang Dong and Ian Horrocks. “Contextual Semantic Embeddings for Ontology Subsumption Prediction”. World Wide Web Journal (WWWJ'23) (Chen et al., 2022).

[7] **Yuan He**, Jiaoyan Chen, Ernesto Jiménez-Ruiz, Hang Dong, and Ian Horrocks.“Language Model Analysis for Ontology Subsumption Inference”. Findings of the Association for Computational Linguistics (ACL'23) (He et al., 2023).

[8] Hang Dong, Jiaoyan Chen, **Yuan He**, Yinan Liu, and Ian Horrocks. “Reveal the Unknown: Out-of-Knowledge-Base Mention Discovery with Entity Linking”. Under review at CIKM'23 (Dong et al., 2023).

[2] Equal contributions. Authors are listed in alphabetical order.

[3] Equal contributions. Authors are listed in alphabetical order.

Note that [5] pertains to the report for the Ontology Alignment Evaluation Initiative (OAEI) 2022. In this event, we took the responsibility of organising the Bio-ML track as proposed in [3]. We plan to maintain our involvement in this initiative in the subsequent years.

We will soon submit another system paper to the Semantic Web Journal (SWJ) for the Python package named DeepOnto[4], and then commence on a new project for ontology embeddings (see Section 4).

4 Final Year Research Plan

Time Period	Plan
Apr 2023 — Jun 2023	Writing the system paper about the DeepOnto library, with an aim to submit to the Semantic Web Journal (SWJ).
Jul 2023 — Nov 2023	Completing the project on ontology embeddings using both language models and hyperbolic geometric models, with an aim to submit to a top-tier machine learning conference.
Dec 2023 — Jan 2024	Commencing preparation for thesis writing, gathering relevant papers and materials for the literature review.
Feb 2024 — May 2024	Flexible planning depending on the result of paper submissions. If time permits, further exploration on ontology embeddings will be undertaken.
Jun 2024 onwards	Focusing on writing the thesis and gathering feedback from supervisors and peers.

Table 1: Tentative schedule for the final D.Phil year.

Our research so far has provided compelling evidence of the utility of language models (LMs) in a range of ontology engineering tasks. As we move forward, the focus shifts to enhancing LMs in domains where they typically falter, such as in encapsulating structural and logical information. The primary objective for the final year of the D.Phil program is to develop robust ontology embeddings that encapsulate not only textual semantics but also structural and logical semantics. These enhanced embeddings can then be deployed in various downstream ontology engineering tasks. This project is expected to yield more theoretical contributions. Concurrently, we have been working

[4] GitHub Repository: `https://github.com/KRR-Oxford/DeepOnto`

on a system paper for DeepOnto, which will complement the thesis from a practical perspective. Detailed summaries of all completed and forthcoming projects can be found in their corresponding chapters in Section 5. A tentative timeline for the final year is also presented in Table 1.

In case that our investigation into the application of hyperbolic models in LM-based ontology embeddings does not yield the desired results, we have a contingency plan. This alternative strategy involves the exclusive use of LMs for constructing ontology embeddings, adopting a "text-for-all" approach, which promises to be simpler yet potentially effective.

5 Thesis Outline

A draft thesis outline for the proposed D.Phil thesis is presented as follows:

Chapter I. Introduction

This chapter will address motivations, provide a brief history and describe the impact of the research, as well as enumerate the contributions of the thesis.

Chapter II. Background

This chapter will provide a literature review on ontology engineering, including various essential sub-tasks, and on language models, which will cover traditional, statistical approaches such as n-gram and naive bayes, as well as neural approaches like RNN and transformer models.

Chapter III. Language Model for Ontology Alignment

The initial part of this chapter will introduce BERTMap (He et al., 2022a), an ontology alignment system for class equivalence that is based on **fine-tuning** the masked language model BERT. The BERTSubs (Chen et al., 2022) system will also be discussed as an extension for subsumption matching. The latter part of this chapter will present Bio-ML (He et al., 2022b), which includes datasets and a comprehensive evaluation framework for ontology alignment models, with a particular focus on those based on machine learning.

Chapter IV. Prompt-based Subsumption Inference

This chapter will primarily delve into the work of OntoLAMA (He et al., 2023), where the task of subsumption inference is defined similarly to natural language inference, aiming to probe an LM's comprehension of ontology semantics. Specifically, an ontology concept verbaliser is developed to transform concept expressions into natural language text; these verbalised concepts are then fed into a template with textual prompts as an LM's inputs for classification. The **prompt learning** paradigm, which allows for the efficient extraction of an LM's knowledge without extensive training, is used in contrast to traditional fine-tuning methods.

Chapter V. Enhancing Language Model for Ontology Embeddings through Hyperbolic Geometry

While large LMs like ChatGPT excel at capturing complex distributional patterns among word tokens (Shanahan, 2022), their capacity to infer structural information remains limited without specific training. To augment an LM's ability to understand structural and logical patterns within an ontology, we propose to embed ontology concepts into **hyperbolic space** using models such as the Poincaré ball. This work is currently underway, and we aim to complete it within the year.

Chapter VI. A Comprehensive Ontology Engineering Package

This chapter introduces DeepOnto, a practical coding library that integrates ontology processing and deep learning modules, with a particular emphasis on LMs. DeepOnto implements basic features like accessing and manipulating entities and axioms, as well as more advanced features that can support deep learning-based ontology engineering systems. For instance, DeepOnto provides ontology verbalisation, a crucial requirement for LM-based solutions, and supports ontology normalisation, an essential component for many logical embedding models. The architecture of DeepOnto is strategically straightforward, comprising essential ontology processing modules and a variety of specific tools and resources built on top of them. This structure ensures an intuitive yet robust framework. Notably, most of the work presented in previous chapters has been integrated into DeepOnto.

Chapter VII. Related Work: Language Model for Other Knowledge Engineering Tasks

This chapter will briefly introduce works related to other knowledge engineering tasks (not restricted to ontology), including those completed during the D.Phil program.

Chapter VII. Conclusion and Discussion

This final chapter will provide a comprehensive summary of the key findings and contributions presented throughout the thesis. We will also engage in thoughtful discussions concerning the broader implications of this research within the field of ontology engineering. Furthermore, we will outline potential avenues for future research. A particularly promising direction lies in investigating the use of large-scale language models for the purpose of knowledge base construction.

References

Ralph Abboud, Ismail Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. 2020. Boxe: A box embedding model for knowledge base completion. *Advances in Neural Information Processing Systems*, 33:9649–9661.

Ryan Brate, Minh-Hoang Dang, Fabian Hoppe, Yuan He, Albert Meroño-Peñuela, and Vijay Sadashivaiah. 2022. Improving language model predictions via prompts enriched with knowledge graphs. In *Workshop on Deep Learning for Knowledge Graphs (DL4KG@ ISWC2022)*.

J Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra AT Moxon, Justin T Reese, Melissa A Haendel, et al. 2023. Structured prompt interrogation and recursive extraction of semantics (spires): A method for populating knowledge bases using zero-shot learning. *arXiv preprint arXiv:2304.02711*.

Jiaoyan Chen, Yuan He, Yuxia Geng, Ernesto Jimenez-Ruiz, Hang Dong, and Ian Horrocks. 2022. Contextual semantic embeddings for ontology subsumption prediction. *arXiv preprint arXiv:2202.09791*.

Hang Dong, Jiaoyan Chen, Yuan He, Yinan Liu, and Ian Horrocks. 2023. Reveal the unknown: Out-of-knowledge-base mention discovery with entity linking. *arXiv preprint arXiv:2302.07189*.

Yuan He, Jiaoyan Chen, Denvar Antonyrajah, and Ian Horrocks. 2021. Biomedical ontology alignment with bert. In *OM@ISWC*.

References

Yuan He, Jiaoyan Chen, Denvar Antonyrajah, and Ian Horrocks. 2022a. Bertmap: A bert-based ontology alignment system. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(5):5684–5691.

Yuan He, Jiaoyan Chen, Hang Dong, Ernesto Jiménez-Ruiz, Ali Hadian, and Ian Horrocks. 2022b. Machine learning-friendly biomedical datasets for equivalence and subsumption ontology matching. In *The Semantic Web–ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings*, pages 575–591. Springer.

Yuan He, Jiaoyan Chen, Ernesto Jiménez-Ruiz, Hang Dong, and Ian Horrocks. 2023. Language model analysis for ontology subsumption inference. *arXiv preprint arXiv:2302.06761*.

Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. 2011. Logmap: Logic-based and scalable ontology matching. In *The Semantic Web–ISWC 2011: 10th International Semantic Web Conference, Bonn, Germany, October 23-27, 2011, Proceedings, Part I 10*, pages 273–288. Springer.

Qiuhao Lu, Nisansa De Silva, Sabin Kafle, Jiazhen Cao, Dejing Dou, Thien Huu Nguyen, Prithviraj Sen, Brent Hailpern, Berthold Reinwald, and Yunyao Li. 2019. Learning electronic health records through hyperbolic embedding of medical ontologies. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 338–346.

Mina Abd Nikooie Pour, Alsayed Algergawy, Patrice Buche, Leyla Jael Castro, Jiaoyan Chen, Hang Dong, Omaima Fallatah, Daniel Faria, Irini Fundulaki, Sven Hertling, Yuan He, Ian Horrocks, Martin Huschka, Liliana Ibanescu, Ernesto Jiménez-Ruiz, Naouel Karam, Amir Laadhar, Patrick Lambrix, Huanyu Li, Ying Li, Franck Michel, Engy Nasr, Heiko Paulheim, Catia Pesquita, Tzanina Saveta, Pavel Shvaiko, Cássia Trojahn, Chantelle Verhey, Mingfang Wu, Beyza Yaman, Ondrej Zamazal, and Lu Zhou. 2022. Results of the ontology alignment evaluation initiative 2022. In *OM@ISWC*, volume 3324 of *CEUR Workshop Proceedings*, pages 84–128. CEUR-WS.org.

Murray Shanahan. 2022. Talking about large language models. *arXiv preprint arXiv:2212.03551*.

Steffen Staab and Rudi Studer. 2010. *Handbook on ontologies*. Springer Science & Business Media.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.