Language Models as Ontology Encoders

Hui Yang[1](✉), Jiaoyan Chen[1], Yuan He[2,4], Yongsheng Gao[3], and Ian Horrocks[4]

[1] The University of Manchester, Manchester, UK
{hui.yang-2,jiaoyan.chen}@manchester.ac.uk
[2] Amazon, Palo Alto, California, USA
lawhy@amazon.com
[3] SNOMED International, London, UK
yga@snomed.org
[4] University of Oxford, Oxford, UK
Ian.Horrocks@cs.ox.ac.uk

Abstract. OWL (Web Ontology Language) ontologies which are able to formally represent complex knowledge and support semantic reasoning have been widely adopted across various domains such as healthcare and bioinformatics. Recently, ontology embeddings have gained wide attention due to their potential to infer plausible new knowledge and approximate complex reasoning. However, existing methods face notable limitations: geometric model-based embeddings typically overlook valuable textual information, resulting in suboptimal performance, while the approaches that incorporate text, which are often based on language models, fail to preserve the logical structure. In this work, we propose a new ontology embedding method OnT, which tunes a Pretrained Language Model (PLM) via geometric modeling in a hyperbolic space for effectively incorporating textual labels and simultaneously preserving class hierarchies and other logical relationships of Description Logic $\mathcal{EL}$. Extensive experiments on four real-world ontologies show that OnT consistently outperforms the baselines including the state-of-the-art across both tasks of prediction and inference of axioms. OnT also demonstrates strong potential in real-world applications, indicated by its robust transfer learning abilities and effectiveness in real cases of discovering new axioms in SNOMED CT construction. Data and code are available at https://github.com/HuiYang1997/OnT.

Keywords: Ontology Embedding · Language Models · Description Logic · Web Ontology Language · Hyperbolic Space

1 Introduction

Ontologies of Web Ontology Language (OWL) can represent explicit, formal, and shared knowledge of a domain, supporting complex knowledge by incorpo-

This work is funded by the EPSRC projects OntoEm (EP/Y017706/1) and ConCur (EP/V050869/1).

Y. He—Work done prior to joining Amazon.

D. Garijo et al. (Eds.): ISWC 2025, LNCS 16140, pp. 443–461, 2026.
https://doi.org/10.1007/978-3-032-09527-5_24

rating Description Logic (DL) axioms [5,11]. These ontologies have become indispensable in domains requiring precise semantic representations; typical examples include the Gene Ontology (GO) [2] in bioinformatics and SNOMED CT [10] in healthcare. With the emergence of neural representation learning techniques [6], there has been growing interests in developing embedding approaches for ontologies that can encode their entities (which include concepts, roles and instances) as numerical vectors while effectively preserving their structural and semantic properties within the vector space for supporting different downstream tasks of prediction, (approximate) inference, retrieval and so on, usually in combination with other machine learning and statistical methods [9,20].

Despite significant advancements, the current methods—which can be divided into two types: *geometric model-based* and *language model-based*—still have distinct shortcomings.

1. *Geometric Model-Based Methods:* These methods represent ontology entities as geometric objects, such as instances as points and concepts as areas, to construct a geometric model of the target ontology [9]. For example, the early method ELEM [19] represents concepts as balls, while more recent methods like BoxEL [33], Box^{2}EL [18], and TransBox [35] represents concepts as boxes. Geometric model-based methods preserve logical relationships by translating DL operators into geometric operations—such as representing concept subsumption as area inclusion and conjunction as intersection—thereby supporting reasoning in the vector space. However, they mostly neglect valuable textual information, such as entity labels that are common in real-world ontologies. This results in suboptimal performance in ontology learning tasks such as axiom prediction, and an inability to embed new entities that are unseen during training—a critical limitation when dealing with dynamic and transfer scenarios.
2. *Language-Model-Based Methods:* These methods, exemplified by OPA2Vec [29] and OWL2Vec* [8], focus on encoding the textual information of ontologies, often following a pipeline which first transforms the axioms and the graph structure into sentences and then tunes a language model to learn entity representations from the sentences [9]. They incorporate both text and formal semantics in the embeddings, which can lead to higher similarities between more related entities [20], but ignore the preservation of logical relationships, which limits their effectiveness in inference. Moreover, most methods generate ontology embeddings using traditional non-contextual word embedding models like Word2Vec, with limited exploration towards the more recent Transformer-based PLMs, which produce layer-specific contextual embeddings rather than general representations. Recently, HiT [17] has been proposed to bridge this gap by training a PLM with geometric modeling for embedding both concept hierarchies and labels. However, HiT is designed for taxonomies, and does not support complex concepts and logical relationships beyond concept subsumption in OWL ontologies.

To address these limitations, we propose **On**tology **T**ransformer encoder (OnT), which integrates the strengths of PLM for contextual text embedding, and geometric modeling in a hyperbolic space for logical structure embedding. OnT enables the preservation of logical relationships of Description Logic $\mathcal{EL}$, thus augmenting axiom inference in the vector space. It effectively incorporates more kinds of semantics for better performance in axiom prediction, and supports the embedding of new entities.

OnT mainly consists of two steps: (1) Complex concepts (denoted as C, D) and roles (denoted by r) are embedded into vector representations using a PLM. The embeddings of complex concepts are derived from a verbalization process that generates textual descriptions for these concepts, while roles are represented as transition functions operating within the space of concept vectors. (2) General Concept Inclusion (GCI) axioms of the form $C \sqsubseteq D$ are represented by regarding them as a hierarchical pre-order $C \prec D$, which is then encoded in a Poincaré ball. Moreover, to effectively capture the logical patterns associated with the existential qualifier (i.e., $\exists r.$) and conjunction (i.e., $\sqcap$), OnT incorporates two specialized loss functions that leverage role embeddings in conjunction with concept embeddings.

Through extensive experiments on real-world ontologies of GALEN [28], Gene Ontology (GO) [2], and Anatomy (Uberon) [25], we demonstrate that our method OnT outperforms current state-of-the-art geometric-model or language-model based approaches in both prediction and inference tasks. Notably, in terms of the Mean Rank metric, OnT achieves up to a sevenfold improvement over existing methods, as observed in the prediction task on the GO dataset. Moreover, it exhibits strong transfer learning capabilities, successfully identifying missing and incorrect direct subsumptions in the SNOMED use case, which highlights the practical potential of our approach for real-world ontology applications.

2 Related Work

Geometric model-based methods encode ontologies by representing their concepts and instances as geometric objects in vector spaces and their roles (i.e., binary relations) as specific geometric relationships between these objects. These methods construct an (approximate) geometric model of the ontology, interpreting logical relationships as geometric meanings. For example, the subsumption of concepts can be understood as the set-inclusion of corresponding geometric objects. Various geometric representations have been explored for representing concepts, including boxes (TransBox [35], Box^{2}EL [18], BoxEL [33], ELBE [27]), balls (ELEM [19], EMEM++ [24]), cones [13,37], and fuzzy sets [30]. The most common way of representing relations is using transition functions defined by the addition of a given vector. Among these approaches, box-based methods have gained prominence due to their closure under intersection—the intersection of two boxes yields another box—enabling them to naturally capture concept conjunctions through geometric operations. In contrast, other geometric representations lack this crucial property, limiting their expressiveness for certain logical operations. The majority of existing methods focus on $\mathcal{EL}$-family ontologies, with notable exceptions of catE [37] and FALCON [30], which provide embeddings for $\mathcal{ALC}$-ontologies.

Language model-based methods originated from early approaches utilizing word embeddings like Word2Vec (which are widely regarded as a kind of neural language models), such as OPA2Vec [29] and OWL2Vec* [8]. They generate embeddings for ontology entities by fine-tuning a word embedding model with the ontology's information, and then apply these embeddings to downstream tasks via an additional, separated prediction model such as a binary classifier. More recently, inspired by the rapid advancement of PLMs based on Transformer architectures [31], a variety of PLM-based approaches such as SORBET and BERTSubs [1,7,14,15,23] have been developed for ontology-related tasks, particularly in the context of ontology completion and alignment. However, these methods jointly fine-tune a PLM and an additional layer that is specific to a downstream task. Thus they do not yield general embeddings that are applicable across different tasks. Furthermore, all language model-based methods—whether based on Word2Vec or transformers—fail to capture logical structures such as the transitivity of subsumption relationships, thereby preventing direct inference within the vector space.

Recently, He et al. proposed HiT [17]—a method that combines language models with hierarchical embedding techniques in hyperbolic spaces to embed taxonomies that consist of hierarchical structures of named concepts. However, this approach overlooks role embeddings and the logical operations that construct complex concepts from basic ones, which are prevalent in real-world ontologies. In this work, we address this limitation with role embeddings and specialized loss functions that capture the logical operators used to build complex concepts from fundamental ones.

We exclude work on Knowledge Graphs such as KG-BERT [36] and KEPLER [32], as our focus is on OWL ontologies, which use Description Logic to model conceptual knowledge—fundamentally different from relational fact-based Knowledge Graphs.

3 Preliminary

3.1 Ontology

OWL ontologies employ sets of statements, known as axioms, to represent and reason about concepts (unary predicates) and roles (binary predicates). In this work, we focus on $\mathcal{EL}$-ontologies, which are investigated by most existing geometric embedding methods. These ontologies strike a balance between expressivity and reasoning efficiency, making them widely applicable [4]. Consider the disjoint sets $\mathsf{N_C} = \{A, B, \ldots\}$, $\mathsf{N_R} = \{r, t, \ldots\}$, and $\mathsf{N_I} = \{a, b, \ldots\}$, representing *concept names* (a.k.a. *atomic or named concepts*), *role names*, and *individual names*, respectively. *$\mathcal{EL}$-concepts* (complex concepts) are defined recursively from these elements as $\top \mid \bot \mid A \mid C \sqcap D \mid \exists r.C \mid \{a\}$. An *$\mathcal{EL}$-ontology* is a finite collection of TBox axioms like $C \sqsubseteq D$, and ABox axioms like $A(a)$ and $r(a, b)$. Note that, through the paper, we denote by atomic concepts as A, B, and any $\mathcal{EL}$-concepts as C, D.

Example 1. Given atomic concepts *Teacher*, *Student*, *Class*, roles *teach*, *hasClass* and *study*, and individuals *Dr.Smith*, *Emma*, there is a small $\mathcal{EL}$-ontology composed of TBox axioms: $\mathit{Person} \sqcap \exists \mathit{teach}.\mathit{Class} \sqsubseteq \mathit{Teacher}$, $\mathit{Person} \sqcap \exists \mathit{study}.\mathit{Class} \sqsubseteq \mathit{Student}$, and ABox axioms: $\mathit{Teacher}(\mathit{Dr.Smith})$, $\mathit{hasClass}(\mathit{Emma}, \mathit{Math101})$.

Normalization of $\mathcal{EL}$-Ontology. In this work, we focus on the TBox part. Note that Abox axioms can be transformed into equivalent TBox axioms by treating instances as classes [18]. An $\mathcal{EL}$-ontology $\mathcal{O}$ is normalized if all its (TBox) axioms are of one of the following forms:

$$A \sqsubseteq B, \quad A_1 \sqcap A_2 \sqsubseteq B, \quad A \sqsubseteq \exists r.B, \quad \exists r.B \sqsubseteq A. \tag{1}$$

For simplicity, we refer to these four types of normalized axioms as NF1-NF4 (where NF denotes normalized form), respectively. It is worth noting that most existing geometric embedding methods are exclusively applicable to normalized ontologies. Any $\mathcal{EL}$-ontology can be transformed into a set of normalized axioms [3] by introducing new atomic concepts along with corresponding names, as illustrated in the following example.

Example 2. To normalize the axiom $\mathit{Person} \sqcap \exists \mathit{teach}.\mathit{Class} \sqsubseteq \mathit{Teacher}$ from Example 1, we introduce a new atomic concept $N_1 \equiv \exists \mathit{teach}.\mathit{Class}$. This transfers the original axiom into three normalized axioms:

$$\mathit{Person} \sqcap N_1 \sqsubseteq \mathit{Teacher}, \quad N_1 \sqsubseteq \exists \mathit{teach}.\mathit{Class}, \quad \text{and} \quad \exists \mathit{teach}.\mathit{Class} \sqsubseteq N_1.$$

Here, the newly introduced concept N_1, derived from $\exists \mathit{teach}.\mathit{Class}$, can be informally interpreted as "Something that teaches some Class."

Inference An *interpretation* $\mathcal{I} = (\Delta^{\mathcal{I}}, \cdot^{\mathcal{I}})$ comprises a non-empty set $\Delta^{\mathcal{I}}$ and a function $\cdot^{\mathcal{I}}$ that maps each $A \in \mathsf{N_C}$ to $A^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}}$, each $r \in \mathsf{N_R}$ to $r^{\mathcal{I}} \subseteq \Delta^{\mathcal{I}} \times \Delta^{\mathcal{I}}$, and each $a \in \mathsf{N_I}$ to $a^{\mathcal{I}} \in \Delta^{\mathcal{I}}$, with $\bot^{\mathcal{I}} = \emptyset$, $\top^{\mathcal{I}} = \Delta^{\mathcal{I}}$, and $\{a\}^{\mathcal{I}} = a^{\mathcal{I}}$. This function extends to any $\mathcal{EL}^{++}$-concepts as follows:

$$(C \sqcap D)^{\mathcal{I}} = C^{\mathcal{I}} \cap D^{\mathcal{I}}, \quad (\exists r.C)^{\mathcal{I}} = \left\{a \in \Delta^{\mathcal{I}} \mid \exists b \in C^{\mathcal{I}} : (a, b) \in r^{\mathcal{I}}\right\},$$

An interpretation $\mathcal{I}$ *satisfies* a TBox axiom $X \sqsubseteq Y$ if $X^{\mathcal{I}} \subseteq Y^{\mathcal{I}}$ for X, Y being two concepts or two role names, or X being a role chain and Y being a role name. It satisfies an ABox axiom $A(a)$ if $a^{\mathcal{I}} \in A^{\mathcal{I}}$ and it satisfies $r(a, b)$ if $(a^{\mathcal{I}}, b^{\mathcal{I}}) \in r^{\mathcal{I}}$. Finally, $\mathcal{I}$ is a *model* of $\mathcal{O}$ if it satisfies every axiom in $\mathcal{O}$. An ontology $\mathcal{O}$ *entails* an axiom α, denoted $\mathcal{O} \models \alpha$, if α is satisfied by all models of $\mathcal{O}$.

3.2 Hyperbolic Space

A d-dimensional *manifold* [21], denoted $\mathcal{M}$, can be regarded as a hypersurface embedded in a higher n-dimensional Euclidean space $\mathbb{R}^n$, which is locally equivalent to $\mathbb{R}^d$ around each point $\mathbf{x} \in \mathcal{M}$. A *Riemannian manifold* $\mathcal{M}$ is a manifold equipped with a Riemannian metric, enabling the definition of a distance function $d_{\mathcal{M}}(\mathbf{x}, \mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in \mathcal{M}$. *Hyperbolic space*, denoted $\mathbb{H}^n$, is a Riemannian manifold with a constant negative curvature of $-\kappa$ ($\kappa > 0$) [22], which can be represented by the Poincaré ball model whose points are defined by a "ball" with radius $1/\sqrt{\kappa}$: $B^n = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| < 1/\sqrt{\kappa}\}$, and the hyperbolic distance between $\mathbf{x}, \mathbf{y} \in B^n$ are defined as

$$d_\kappa(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{\kappa}} \operatorname{arcosh}\left(1 + \frac{2\kappa\|\mathbf{x} - \mathbf{y}\|^2}{(1 - \kappa\|\mathbf{x}\|^2)(1 - \kappa\|\mathbf{y}\|^2)}\right). \tag{2}$$

In the Poincaré ball model, the scaling $k \in R$ of a point $\mathbf{x} \in B^n$ is defined as

$$k \odot \mathbf{x} = \tanh(k \cdot \tanh^{-1}(||\mathbf{x}||)) \cdot \frac{\mathbf{x}}{||\mathbf{x}||} \tag{3}$$

4 Methodology

In this section, we present our method, OnT, for embedding a given $\mathcal{EL}$-ontology. OnT consists mainly of three parts:

1. Embedding any $\mathcal{EL}$-concepts (atomic or complex ones) as points in hyperbolic spaces using PLMs and verbalizations (Sect. 4.1);
2. Embedding roles as rotations over hyperbolic spaces (Sect. 4.2), which allows OnT for capturing the logical structures of existential qualifications $\exists r$ (Proposition 1) and demonstrates improved performance as evidenced by evaluations on real-world ontologies;
3. Training the embeddings using the Poincaré ball model (Sect. 4.3) by regarding the axioms as hierarchical relationships between complex concepts.

It is worth noting that the role embedding component can be omitted to yield a simplified variant, referred to as OnT(w/o r).

4.1 Verbalisation-Based Concept Embedding

Given an ontology $\mathcal{O}$, we assume that each atomic concept A and role r occurring in $\mathcal{O}$ is associated with a textual description, typically its name or definition, denoted as $\mathcal{V}(A)$ and $\mathcal{V}(r)$, respectively. For instance, we may have $\mathcal{V}(A) =$ "Father" and $\mathcal{V}(r) =$ "is parent of".

Based on these descriptions of atomic concepts and roles, we systematically generate a natural language description for each complex concept C appearing in the ontology $\mathcal{O}$, denoted as $\mathcal{V}(C)$. For $\mathcal{EL}$-ontologies, we generate these descriptions according to the following compositional rules:

$$\mathcal{V}(C \sqcap D) = \text{“}\mathcal{V}(C) \text{ and } \mathcal{V}(D)\text{”}, \quad \mathcal{V}(\exists r.C) = \text{“ something that } \mathcal{V}(r) \text{ some } \mathcal{V}(C)\text{”}.$$

For example, we will have $\mathcal{V}(\text{Person} \sqcap \text{Student}) =$ "person and student", and $\mathcal{V}(\exists \text{isParentOf.Person}) =$ "something that is parent of some person".

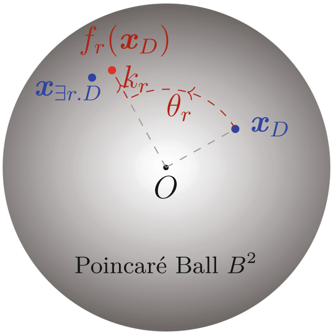


Fig. 1. Illustration of f_r in a two-dimensional hyperbolic space.

With the verbalization approach described above, we can embed any complex $\mathcal{EL}$-concept C by applying language models to its textual description $\mathcal{V}(C)$ and mapping the result to a point in hyperbolic space as in HiT [17]. Specifically, this is achieved by encoding sentences using a BERT model with mean pooling, after which the resulting embeddings are re-trained in hyperbolic space. The final embedding of C is denoted as $\boldsymbol{x}_C$.

4.2 Logic-Aware Role Embedding

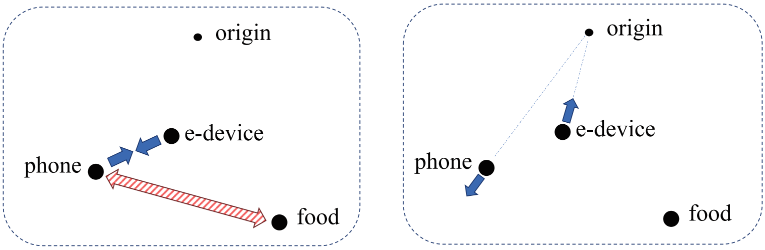
In the above verbalization process, the role semantics is integrated into the concept verbalizations. However, this do not provide individual roles embeddings, which could restrict the capability of handling logical patterns involving roles and impairing the reasoning. For instance, we can not guarantee the preservation of deductive patterns such as the monotonicity of existential restrictions: if $A \sqsubseteq B$, then $\exists r.A \sqsubseteq \exists r.B$.

To address these limitations, we propose to explicitly incorporate role embeddings by interpreting a role r as a function f_r over hyperbolic space. In details, for each complex concept of the form $\exists r.D$, OnT introduces an alternative representation: $f_r(\boldsymbol{x}_D)$, which complements the verbalization-based embedding $\boldsymbol{x}_{\exists r.D}$. Here, f_r is a role-specific transformation function. We will encourage the two embeddings $f_r(\boldsymbol{x}_D)$ and $\boldsymbol{x}_{\exists r.D}$ to be identical by introducing an extra loss term in the training process.

In our implementation, we define f_r as a composition of rotations and scaling operations in hyperbolic space. Specifically, the f_r is defined by (see Fig. 1 for an illustration):

$$f_r(\mathbf{v}) = k_r \odot (R(\Theta_r) \cdot \mathbf{v}), \tag{4}$$

where $k_r \in \mathbb{R}$ is a role-specific scaling factor, $\Theta_r = (\theta_r^1, \theta_r^2, \ldots, \theta_r^m) \in \mathbb{R}^m$ is a role-specific rotation angle, and $\mathbf{v} \in \mathbb{H}^{2m}$ is a point in hyperbolic space. Here, the $\odot$ operation represents the scaling product over hyperbolic space $\mathbb{H}^{2m}$, which ensures the scaled embeddings are still in hyperbolic space. However, for rotations, we could directly apply the same rotations as the Euclidean space as we



(a) Hierarchical Contrastive Loss (b) Centripetal Loss

Fig. 2. Illustration of impact of hierarchy Loss $\mathcal{L}_{\prec}$ during training.

use the Poincaré ball models for the representation of hyperbolic space. Specifically, we use the rotation matrix $R(\Theta_r) \in \mathbb{R}^{2m\times 2m}$ over the space $\mathbb{H}^{2m}$ defined as a product of two-dimensional rotations of the following form:

$$R(\Theta_r) = \begin{bmatrix} R(\theta_r^1) & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & R(\theta_r^2) & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & R(\theta_r^m) \end{bmatrix}, \text{ where } R(\theta_r) = \begin{bmatrix} \cos(\theta_r) & -\sin(\theta_r) \\ \sin(\theta_r) & \cos(\theta_r) \end{bmatrix}.$$

4.3 Training

Hierarchy Loss. We interpret subsumption axioms $C \sqsubseteq D$ in the ontology $\mathcal{O}$ as partial-order relationships between their embeddings: $\boldsymbol{x}_C \prec \boldsymbol{x}_D$. Then, following the approach of HiT [17], we encode these partial-order relationships using a Poincaré embedding model [26] using a hierarchical loss defined by the hyperbolic distance. The loss function $\mathcal{L}_{\prec}(\boldsymbol{x}_C \prec \boldsymbol{x}_D)$ consists of two parts:

1. *Hierarchical Contrastive Loss*: This loss encourages embeddings of related concepts to be closer to each other than to negative samples:

$$\mathcal{L}_{contrast}(\boldsymbol{x}_C \prec \boldsymbol{x}_D) = \max(0, d_\kappa(\boldsymbol{x}_C, \boldsymbol{x}_D) - d_\kappa(\boldsymbol{x}_C, \boldsymbol{x}_{D_{\text{neg}}}) + \alpha),$$

   where D_{neg} represents a randomly sampled concept that composes a negative example with C and α is a margin hyperparameter.
2. *Centripetal Loss*: This loss enforces that parent concepts are embedded closer to the origin than their children in the hyperbolic space. Let $\|\boldsymbol{x}\|_\kappa$ denote the hyperbolic distance from a point $\boldsymbol{x} \in \mathbb{H}^n$ to the origin (also known as the hyperbolic norm). The centripetal loss is defined as:

$$\mathcal{L}_{centri}(\boldsymbol{x}_C \prec \boldsymbol{x}_D) = \max(0, \|\boldsymbol{x}_D\|_\kappa - \|\boldsymbol{x}_C\|_\kappa + \beta),$$

   where β is a margin hyperparameter. This constraint geometrically reinforces the hierarchical structure by positioning more general concepts (parents) closer to the center of the hyperbolic space.

The overall hierarchy loss is defined as the sum of these two loss components:

$$\mathcal{L}_{\prec}(\boldsymbol{x}_C \prec \boldsymbol{x}_D) = \mathcal{L}_{contrast}(\boldsymbol{x}_C \prec \boldsymbol{x}_D) + \mathcal{L}_{centri}(\boldsymbol{x}_C \prec \boldsymbol{x}_D). \tag{5}$$

The effect of the hierarchy loss on embedding updates is illustrated in Fig. 2, where the positive pair is $C =$ *"phone"* and $D =$ *"e-device"*, and the negative example is $D_{\text{neg}} =$ *"food"*. The contrastive loss encourages the embeddings of C and D to be close, while pushing C and D_{neg} farther apart, as shown in Fig. 2a. On the other hand, in in Fig. 2b, the centripetal loss pulls the parent concept D toward the origin, while pushing the child concept C away from it.

Loss for Role Embeddings. The loss for role embeddings aims to align the embeddings of $\boldsymbol{x}_{\exists r.D}$ with $f_r(\boldsymbol{x}_D)$. However, as shown by our preliminary experiments, it is not a good choice to directly align the embeddings such as introducing a loss defined by their Euclidean distance or hyperbolic, i.e., $||\boldsymbol{x}_{\exists r.D} - f_r(\boldsymbol{x}_D)||$ or $d_\kappa(\boldsymbol{x}_{\exists r.D}, f_r(\boldsymbol{x}_D))$. Instead, we would reuse the hierarchical loss above by interpreting the equivalence $\boldsymbol{x}_{\exists r.D} \equiv f_r(\boldsymbol{x}_D)$ as two partial-order $\boldsymbol{x}_{\exists r.D} \prec f_r(\boldsymbol{x}_D)$ and $f_r(\boldsymbol{x}_D) \prec \boldsymbol{x}_{\exists r.D}$. Formally, the loss is defined as:

$$\mathcal{L}_r(\exists r.D) = \frac{1}{2}\Big(\mathcal{L}_{\prec}\big(\boldsymbol{x}_{\exists r.D} \prec f_r(\boldsymbol{x}_D)\big) + \mathcal{L}_{\prec}\big(f_r(\boldsymbol{x}_D) \prec \boldsymbol{x}_{\exists r.D}\big)\Big) \tag{6}$$

Loss for Conjunction. This loss is introduced to capture the logical properties of the conjunction $\sqcap$, specifically, a universally valid axiom $C \sqcap D \sqsubseteq C$. It is enough to use the following loss based on the hierarchy loss $\mathcal{L}_{\prec}$:

$$\mathcal{L}_{\sqcap}(C \sqcap D) = \frac{1}{2}\Big(\mathcal{L}_{\prec}(\boldsymbol{x}_{C\sqcap D} \prec \boldsymbol{x}_C) + \mathcal{L}_{\prec}(\boldsymbol{x}_{C\sqcap D} \prec \boldsymbol{x}_D)\Big). \tag{7}$$

Training. The final train loss is defined as the sum of the losses defined by Eqs. (5), (6), and (7) for all axioms $C \sqsubseteq D$, concept $\exists r.D$, and conjunctions $C \sqcap D$ appeared $\mathcal{O}$, respectively.

Finally, with a well-trained OnT model, we evaluate a new axiom $C \sqsubseteq D$ using the following score with a higher value indicates a higher confidence of the given axioms, which is defined as a weighted sum of distances:

$$s(C \sqsubseteq D) \equiv s(\boldsymbol{x}_C \prec \boldsymbol{x}_D) := -(d_\kappa(\boldsymbol{x}_C, \boldsymbol{x}_D) + \lambda(\|\boldsymbol{x}_D\|_\kappa - \|\boldsymbol{x}_C\|_\kappa)) \tag{8}$$

where the weight λ is determined based on the model's performance on the validation set, higher scores indicate a stronger predicted subsumption relationship between concepts.

We have the following proposition that allows us to control the difference between scores $s(f_r(\boldsymbol{x}_C) \prec f_r(\boldsymbol{x}_D))$ and $s(\boldsymbol{x}_C \prec \boldsymbol{x}_D)$ using the scaling factor k_r, and thus, capturing the deductive pattern $A \sqsubseteq B \Rightarrow \exists r.A \sqsubseteq \exists r.B$.

Proposition 1. *For any* $\mathbf{x}, \mathbf{y} \in \mathbb{H}^{2m}$ *and rotation matrix* $R(\Theta_r) \in \mathbb{R}^{2m\times 2m}$ *as defined in Eq.* (4)*, we have* $\|\boldsymbol{x}_C\|_\kappa = \|R(\Theta_r)\cdot \boldsymbol{x}_C\|_\kappa$ *and* $d_\kappa(\mathbf{x}, \mathbf{y}) = d_\kappa(R(\Theta_r)\cdot \mathbf{x}, R(\Theta_r)\cdot \mathbf{y})$*. Moreover, we have* $s(f_r(\boldsymbol{x}_C) \prec f_r(\boldsymbol{x}_D)) = s(\boldsymbol{x}_C \prec \boldsymbol{x}_D)$ *when* $k_r = 1$*.*

Proof. The hyperbolic distance $d_\kappa(\mathbf{x}, \mathbf{y})$ depends only on the Euclidean norms $\|\mathbf{x}\|$ and $\|\mathbf{y}\|$, as per Eq. (2). Since the rotation $R(\Theta_r)$ preserves Euclidean norms, it follows that $\|R(\Theta_r) \cdot \mathbf{z}\| = \|\mathbf{z}\|$ for $\mathbf{z} = \mathbf{x}, \mathbf{y}$. Therefore, $d_\kappa(\mathbf{x}, \mathbf{y}) = d_\kappa(R(\Theta_r) \cdot \mathbf{x}, R(\Theta_r) \cdot \mathbf{y})$. By definition, we have $\|\mathbf{x}\|_\kappa = d_\kappa(\mathbf{x}, \mathbf{0})$. Applying this with $\mathbf{y} = \mathbf{0}$, we obtain: $\|\mathbf{x}\|_\kappa = \|R(\Theta_r) \cdot \mathbf{x}\|_\kappa$.

Since the score is defined by $d_\kappa(\mathbf{x}, \mathbf{y})$, $\|\mathbf{x}\|_\kappa$, and $\|\mathbf{y}\|_\kappa$, and given $f_r(\mathbf{z}) = R(\Theta_r) \cdot \mathbf{z}$ when $k_r = 1$, we conclude that $s(f_r(\boldsymbol{x}_C) \prec f_r(\boldsymbol{x}_D)) = s(\boldsymbol{x}_C \prec \boldsymbol{x}_D)$ when $k_r = 1$. This completes the proof.

5 Evaluation

5.1 Experiment Setting

The evaluation is mainly concentrated on two tasks: axiom prediction (Sect. 5.2) and inference (Sect. 5.3). The prediction and inference tasks focus on identifying missing axioms; however, the prediction task addresses arbitrary axioms, while the inference task focuses specifically on axioms that can be logically derived from the given ontologies. We also evaluate the performance of our method in different scenarios such as transfer learning, ablation study, and over real cases in Sect. 5.4.

Datasets. We adopt three real-world ontologies—GALEN [28], the Gene Ontology (GO) [2], and Anatomy (Uberon) [25]. Following the prior research [18,35], we keep only the $\mathcal{EL}$ part and use their normalized versions. For the prediction task, the training, validation, and testing data are generated by a random 80/10/10 split of the ontology axioms. For the inference task, we use the whole ontology as the training data, and all the inferred axioms of NF1 as the testing data, and 1000 randomly selected inferred NF1 subsumptions as validation data. The data statistics are shown in Table 1. Note that we developed our own ontology normalization implementation rather than using the existing implementation in ELEM [19], mOWL [38], and DeepOnto [16] as it (1) does not name the concepts introduced during normalization, and (2) sometimes produces logically inconsistent axioms[1] due to some bug in calling the jcel normalizer.

Baselines. Our study systematically compares our proposed methods with established approaches that provide general ontology embeddings, with particular emphasis on geometric embedding methods, including Box^{2}EL [18], BoxEL [33], TransBox [35], ELBE [27], and ELEM [19]. We exclude catE and FALCON as catE cannot handle unseen complex concepts that appear in our experimental settings, and FALCON's implementation is not publicly available. Additionally, we benchmark against HiT [17], a language model-based method limited to taxonomic structures (i.e., NF1 axioms), and two classic none contextual word embedding-based methods—OPA2Vec [29] and OWL2Vec* [8]. We

[1] For example, among the normalized axioms of GALEN, we find the axiom $\exists_{hasQuantity}$ *BNFSection13_3* $\sqsubseteq$ *Tobacco*, which contradicts the original ontology where *BNFSection13_3* appears only in *BNFSection13_3* $\sqsubseteq$ *BNFChapter13Section*.

Table 1. Normalized Dataset Statistics (Train/Val/Test for prediction task).

Axioms	GALEN	GO	ANATOMY
NF1	25,610/3,200/3,203	116,751/14,593/14,596	41,764/5,220/5,222
NF2	11,679/1,459/1,462	24,097/3,011/3,014	12,336/1,542/1,543
NF3	25,299/3,161/3,165	238,899/29,861/29,865	39,766/4,970/4,972
NF4	6,287/785/788	81,948/10,243/10,245	7,586/947/951
Total	68,875/8,605/8,618	461,695/57,708/57,720	101,452/12,679/12,688
Inferred(NF1)	335,002	1,184,380	225,330

also include a simplified version of OnT, denoted as OnT(w/o r), which omits role embeddings and is trained using only the loss of Eq. 5. We ignored other PLM fine-tuning-based methods like BERTSub [7] whose embeddings are coupled to a task-specific layer without generality towards different tasks.

Evaluation Metrics. Consistent with the established literature [18,19,27,33,35], we evaluate ontology embedding performance using various ranking-based metrics on the testing set. We rank candidates according to the score function defined in Eq. 8, where higher scores indicate more probable candidates. To comprehensively assess different methods, we track the rank of correct answers and report performance through several standard metrics: Hits@k (H@k) for $k \in \{1, 10, 100\}$, mean reciprocal rank (MRR), and mean rank (MR).

Experimental Protocol. We mainly use all-MiniLM-L12-v2 (33.4M) as the underlying language model for OnT and HiT. The influence of different language models is presented in Sect. 5.4. We trained OnT and HiT for 1 epoch, and with 1 negative sample for each given axiom, as we found that it would be enough to get good performance in the pre-test. The embedding vectors for each concept is obtained by performing an average pooling over features of the final layer of the language model. For obtaining the vector for $\Theta(r), k_r$ for a given role r, we apply an extra linear transformation on the embedding of r. The margins α, β and learning rate γ are fixed as default in HiT as $3.0, 0.5, 10^{-5}$, respectively. The weight $\lambda \in \{0, 0.1, \ldots, 1\}$ of the score function in Eq. 8 is selected based on the best performance on the validation set.

OWL2Vec* and OPA2Vec utilize fine-tuned word embeddings (https://tinyurl.com/word2vec-model) with the Random Forest classifier for superior performance, except for GO ontology inference tasks, where Logistic Regression is employed due to computational constraints. Due to dataset modifications, we also re-implemented all the other geometric embedding models (BoxEL, TransBox, ELBE and ELEM) based on the framework developed by Box^{2}EL [18] and TransBox [35]. For our implementation, we utilized embedding dimensions $d = 200$, explored margin values $\gamma \in \{0, 0.05, 0.1, 0.15\}$ and learning rates $l_r \in \{0.0005, 0.005, 0.01\}$, and trained each model for 5,000 epochs. Optimal hyperparameters were selected based on validation set performance.

Table 2. Overall performance of the prediction task across datasets. Values for H@k and MRR are percentages. $k = 1/10/100$ for H@k.

	GALEN			GO			ANATOMY		
Method	H@k	MRR	MR	H@k	MRR	MR	H@k	MRR	MR
ELEM	14/**50**/68	26	2,715	4/35/68	14	11,764	10/53/78	24	1,588
ELBE	9/37/55	18	4,661	10/30/43	18	10,236	9/41/66	20	2,672
BoxEL	0/0/2	0	13,824	0/0/2	0	65,846	1/2/4	2	12,257
Box^{2}EL	12/38/58	21	4,593	8/43/64	19	7,975	11/39/65	20	2,828
TransBox	11/41/62	22	2,972	8/43/67	19	7,092	9/49/73	22	1,299
OPA2Vec	0/1/4	1	13,547	0/1/4	0	18,493	0/5/17	2	9,537
OWL2Vec*	0/1/5	1	13,660	0/0/2	0	19,523	0/3/11	2	10,309
HiT	**25**/47/62	33	2,349	36/60/73	44	15,080	19/54/78	31	722
OnT(w/o r)	**26**/46/64	33	1,546	**38**/66/79	**48**	2,209	**22**/52/79	31	628
OnT	**25**/**50**/**69**	**34**	**792**	37/**67**/**81**	46	**1,121**	**22**/**57**/**82**	**33**	**475**

5.2 Prediction Task

The comprehensive evaluation results are presented in Table 2 for GALEN, GO, and Anatomy, respectively. Our method OnT consistently outperforms existing approaches across all datasets. While some geometric model-based methods achieve comparable performance in terms of H@k, such as ELEM in the GALEN dataset, they typically exhibit substantially lower average performance, which is evidenced by the significant gap between MRR and MR values. For instance, the best-performing geometric-based method on GO, Transbox, yielded MR values approximately 7 times worse and MRR values twice as poor as OnT. This indicates that OnT has overall fewer extreme worst cases (i.e., correct answers with extremely large ranks) and also better cases (i.e., lower rankings). This phenomenon occurs consistently across all three ontologies.

For language model-based methods, we observe that OPA2Vec and OWL2Vec* demonstrate limited performance, which is reasonable given their reliance on word embeddings and random forest classifiers for capturing subsumptions. It is important to note that, in our evaluation, ranking is performed over all atomic concepts as candidates—unlike the original OWL2Vec settings [8], which consider only around 50 candidates—making metrics such as Hits@k not directly comparable. In contrast, by employing more advanced BERT-based language models and geometric embeddings based on hyperbolic spaces, HiT achieved significantly better performance. By further incorporating the logical constraints of complex concepts, our method OnT outperformed HiT, especially in terms of average performance as indicated by MR values. Specifically, OnT achieved approximately 14 times better MR values than HiT in the GO dataset, suggesting that OnT could more effectively avoid extremely poor cases while also improving performance on other metrics such as H@k and MRR. Moreover, we can see that, in most cases, adding role embeddings and extra loss for logical constraints of ∃ and ⊓ lead to better performance by comparing the OnT(w/o r) and OnT.

Table 3. Performance of the inference task across datasets. Values for H@k and MRR are percentages. $k = 1/10/100$ for H@k.

	GALEN			GO			ANATOMY		
Method	H@k	MRR	MR	H@k	MRR	MR	H@k	MRR	**MR**
ELEM	0/3/9	1	8,639	0/3/17	1	18,377	0/4/22	2	4,990
ELBE	0/4/20	**2**	2,999	0/3/15	1	4,021	0/6/39	2	979
BoxEL	0/0/3	0	11,328	0/0/0	0	18,186	0/0/0	0	8,169
Box^{2}EL	0/3/15	1	5,530	0/1/7	1	11,801	0/1/7	1	11,801
TransBox	0/2/6	1	7,111	0/2/9	1	4,449	0/5/27	2	749
OPA2Vec	0/0/1	0	12,722	3/5/6	3	95,755	**2**/6/15	**3**	5,143
OWL2Vec*	0/0/1	0	12,647	3/7/8	**5**	88,614	1/5/14	**3**	5,441
HiT	0/4/26	**2**	953	0/1/4	0	44,253	0/6/**44**	**3**	**441**
OnT(w/o r)	0/4/20	1	1,047	0/5/39	2	**824**	0/**7**/40	**3**	499
OnT	0/**5**/**28**	**2**	**913**	0/**10**/**40**	3	832	0/6/41	**3**	458

5.3 Inference Task

The overall results are summarized in Table 3. We can see that OnT clearly outperforms existing geometric model-based embedding methods across all datasets. In particular, on the GO dataset, our method achieves approximately 3 times better H@10 and H@100, and 5 times better MR. This improvement may reflect the advantages of hyperbolic embeddings, as HiT also shows strong results in the GALEN and ANATOMY datasets. However, HiT performs poorly on the GO dataset. This suggests that the information encoded in NF2–NF4 axioms, which are more prominent in GO than in other datasets, plays a crucial role. Since HiT does not incorporate this information during training, its performance suffers; such trend has also been reflected in the MR metrics on the prediction task. Furthermore, we observe that in most cases, incorporating role embeddings and losses for logical constraints allows OnT(w/o r) to achieve even better performance than both OnT and HiT.

The overall performance of OPA2Vec and OWL2Vec* is lower than that of most other methods. This is expected, as both OPA2Vec and OWL2Vec* are prediction-based approaches that evaluate axioms using a binary classifier. Such methods struggle to capture complex logical relationships—like the transitivity of SubClassOf relations—which limits their effectiveness in inference tasks.

Table 4. Ablation study for prediction and inference tasks on GALEN.

Task	Method	Language Model	H@1	H@10	H@100	MRR	MR
Prediction	OnT(w/o r)	all-MiniLM-L6-v2	26	47	64	33	1,972
		all-MiniLM-L12-v2	26	46	64	33	1,546
		all-MPNet-base-v2	26	43	63	32	979
	OnT	all-MiniLM-L6-v2	26	**51**	69	35	1,131
		all-MiniLM-L12-v2	25	50	69	34	792
		all-MPNet-base-v2	**27**	50	**72**	**34**	**630**
Inference	OnT(w/o r)	all-MiniLM-L6-v2	0	**5**	23	2	1,062
		all-MiniLM-L12-v2	0	4	20	1	1,047
		all-MPNet-base-v2	0	2	13	1	1,148
	OnT	all-MiniLM-L6-v2	0	3	26	2	923
		all-MiniLM-L12-v2	0	**5**	28	2	913
		all-MPNet-base-v2	0	4	**34**	2	**630**

5.4 Other Results

Ablation Study. To evaluate the impact of different language models and loss functions, we follow the methodology of [17] and experiment with two additional top-performing pre-trained models from the Sentence Transformers library: all-MiniLM-L6-v2 (22.7M parameters) and all-mpnet-base-v2 (109M parameters), in addition to the all-MiniLM-L12-v2 model (33.4M parameters) used in our main experiments. From Table 4, we can see that the performance differences among these models are relatively small. While the larger model consistently shows better average performance, as indicated by improved MR values, it does not always outperform the smaller models across all evaluation metrics.

Transfer Learning. We evaluated the OnT and HiT models in a transfer learning paradigm, where each model was trained and evaluated on a source dataset, then tested on a distinct target dataset, using three different datasets in the prediction task. The overall transfer learning performance of OnT and HiT using MiniLM-L12-v2 is illustrated by Fig. 3. We can see that OnT and HiT both achieve good transfer abilities, while OnT performance better, indicated by the consistently lower MR value, and the higher H@100 or MRR value in most of cases. Especially for the cases from GO to GALEN or ANATOMY.

Case Study. In our case study, we evaluate real-world scenarios encountered during the construction of ontologies, particularly in the development of a new anatomy ontology derived from SNOMED CT. The following two cases, summarized in Fig. 4, illustrate the potential of our model as a valuable tool in ontology construction.

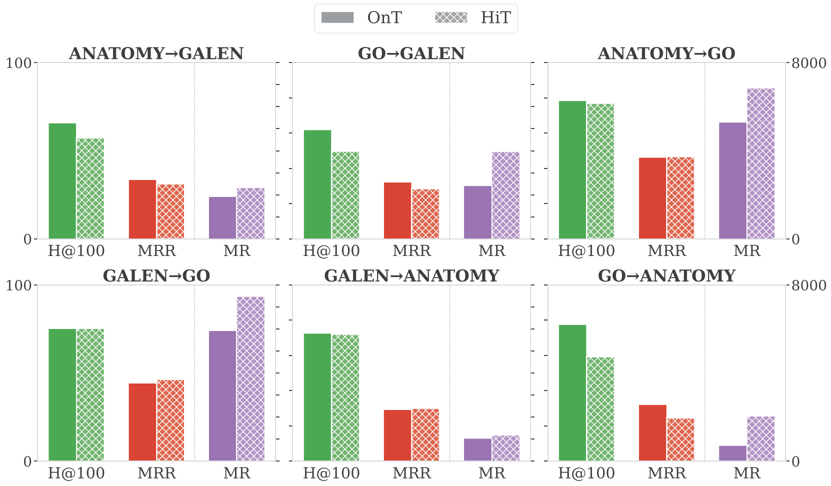


Fig. 3. Transfer learning results of OnT and HiT with MiniLM-L12-v2.

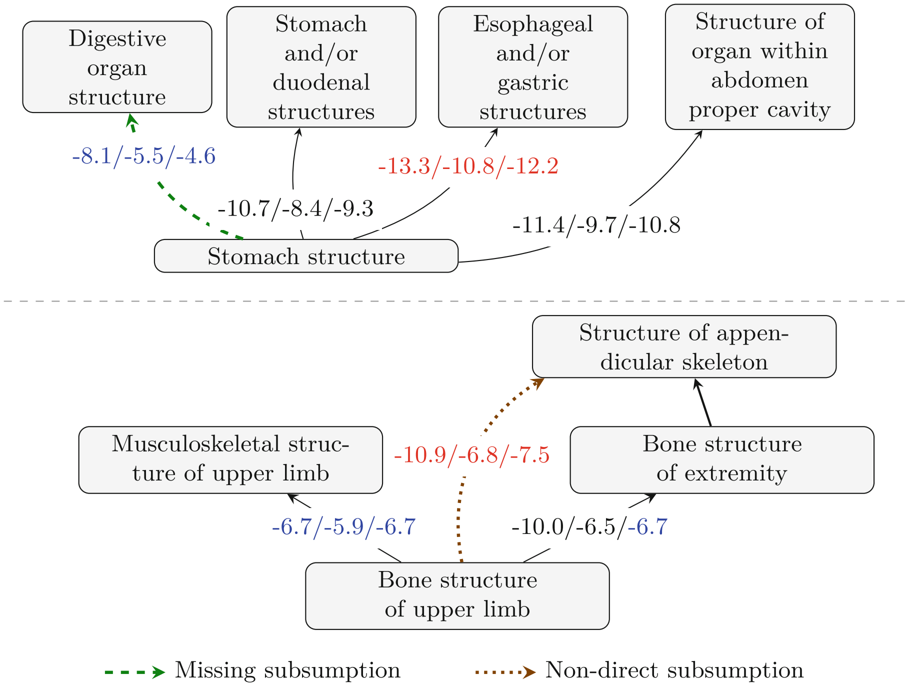

Digestive organ structure
Stomach and/or duodenal structures
Esophageal and/or gastric structures
Structure of organ within abdomen proper cavity
-8.1/-5.5/-4.6
-10.7/-8.4/-9.3
-13.3/-10.8/-12.2
-11.4/-9.7/-10.8
Stomach structure
Structure of appendicular skeleton
Musculoskeletal structure of upper limb
-10.9/-6.8/-7.5
Bone structure of extremity
-6.7/-5.9/-6.7
-10.0/-6.5/-6.7
Bone structure of upper limb
Missing subsumption
Non-direct subsumption

Fig. 4. Case Study: Arrows $C \rightarrow D$ represent subsumption $C \sqsubseteq D$ with scores from Eq. 8. Each arrow shows three scores from three OnT models trained on GALEN/GO/ANATOMY ontologies, respectively. A higher score indicates a more likely subsumption. Blue/red highlights indicate the highest/lowest scores for all subsumptions with the same subclass.

1. *Missing Subsumptions:* In the manually constructed ontology of SNOMED CT, a direct subsumption is overlooked: "Stomach structure $\sqsubseteq$ Digestive organ structure". Our method is proven effective in identifying this missing subsumption, consistently assigning it a higher score than other existing superclasses of the "Stomach structure" within the constructed ontology.
2. *Erroneous (Direct) Subsumptions:* We detect an incorrect direct Superclass of "Bone structure of upper limb" as "Structure of appendicular skeleton", which is incorrect as "Bone structure of extremity" should have such a parent. Our model effectively identifies this erroneous relationship by consistently assigning it the lowest score among all existing superclasses.

6 Conclusion and Future Work

In this study, we introduce OnT, which integrates geometric models with language models to derive ontology embeddings for concepts and roles. Through extensive experiments on real-world ontologies, we demonstrate that our approach achieves state-of-the-art performance in both prediction (inductive reasoning) and inference (deductive reasoning) tasks. Furthermore, our method exhibits strong transfer learning capabilities, suggesting its potential for real-world applications in related domains.

Looking ahead, our future research aims to merge our current methodologies with other hierarchical embedding techniques, such as [12,34]. Additionally, we are keen to extend our methods to more complex ontology languages, such as extending to $\mathcal{ALC}$ with the negation logical operator ¬, or delve deeper into the logical patterns of roles using the role embeddings generated by OnT, including investigating role inclusion axioms. It would also be interesting to conduct a more thorough analysis of our model, such as exploring the impact of verbalization quality, or performance across a wider range of ontologies beyond those currently utilized.

Supplementary Materials. All supplementary materials, including the code and dataset, are available at https://github.com/HuiYang1997/OnT.

References

1. Amir, M., et al.: Truveta mapper: a zero-shot ontology alignment framework. In: Shvaiko, P., Euzenat, J., Jiménez-Ruiz, E., Hassanzadeh, O., Trojahn, C. (eds.) Proceedings of the 18th International Workshop on Ontology Matching co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023. CEUR Workshop Proceedings, vol. 3591, pp. 1–12. CEUR-WS.org (2023). https://ceur-ws.org/Vol-3591/om2023_LTpaper1.pdf
2. Ashburner, M., et al.: Gene ontology: tool for the unification of biology. Nat. Genet. **25**(1), 25–29 (2000)

3. Baader, F., Brandt, S., Lutz, C.: Pushing the EL envelope. In: Kaelbling, L.P., Saffiotti, A. (eds.) IJCAI 2005, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, 30 July - 5 August 2005, pp. 364–369. Professional Book Center (2005), http://ijcai.org/Proceedings/05/Papers/0372.pdf
4. Baader, F., Gil, O.F.: Extending the description logic EL with threshold concepts induced by concept measures. Artif. Intell. **326**, 104034 (2024)
5. Baader, F., Horrocks, I., Sattler, U.: Description logics as ontology languages for the semantic web. In: Hutter, D., Stephan, W. (eds.) Mechanizing Mathematical Reasoning. LNCS (LNAI), vol. 2605, pp. 228–248. Springer, Heidelberg (2005). https://doi.org/10.1007/978-3-540-32254-2_14
6. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
7. Chen, J., He, Y., Geng, Y., Jiménez-Ruiz, E., Dong, H., Horrocks, I.: Contextual semantic embeddings for ontology subsumption prediction. World Wide Web (WWW) **26**(5), 2569–2591 (2023)
8. Chen, J., Hu, P., Jiménez-Ruiz, E., Holter, O.M., Antonyrajah, D., Horrocks, I.: Owl2vec*: embedding of OWL ontologies. Mach. Learn. **110**(7), 1813–1845 (2021)
9. Chen, J., Mashkova, O., Zhapa-Camacho, F., Hoehndorf, R., He, Y., Horrocks, I.: Ontology embedding: a survey of methods, applications and resources. arXiv preprint arXiv:2406.10964 (2024)
10. Donnelly, K., et al.: Snomed-ct: the advanced terminology and coding system for ehealth. Stud. Health Technol. Inform. **121**, 279 (2006)
11. Fitz-Gerald, S.J., Wiggins, B.: Staab, s., studer, r. (eds.), handbook on ontologies, series: International handbooks on information systems, second ed., vol. XIX (2009), 811 p., 121 illus., hardcover £164, ISBN: 978-3-540-70999-2. Int. J. Inf. Manag. **30**(1), 98–100 (2010). https://doi.org/10.1016/J.IJINFOMGT.2009.11.012
12. Ganea, O., Bécigneul, G., Hofmann, T.: Hyperbolic entailment cones for learning hierarchical embeddings. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, 10-15 July 2018. Proceedings of Machine Learning Research, vol. 80, pp. 1632–1641. PMLR (2018). http://proceedings.mlr.press/v80/ganea18a.html
13. Garg, D., Ikbal, S., Srivastava, S.K., Vishwakarma, H., Karanam, H.P., Subramaniam, L.V.: Quantum embedding of knowledge for reasoning. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada, pp. 5595–5605 (2019). https://proceedings.neurips.cc/paper/2019/hash/cb12d7f933e7d102c52231bf62b8a678-Abstract.html
14. Gosselin, F., Zouaq, A.: SORBET: A siamese network for ontology embeddings using a distance-based regression loss and BERT. In: Payne, T.R., et al.(eds.) The Semantic Web - ISWC 2023 - 22nd International Semantic Web Conference, Athens, Greece, November 6-10, 2023, Proceedings, Part I. LNCS, vol. 14265, pp. 561–578. Springer (2023). https://doi.org/10.1007/978-3-031-47240-4_30
15. He, Y., Chen, J., Antonyrajah, D., Horrocks, I.: Bertmap: A bert-based ontology alignment system. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, 22 February - 1 March 2022, pp. 5684–5691. AAAI Press (2022). https://doi.org/10.1609/AAAI.V36I5.20510

16. He, Y., et al.: Deeponto: A python package for ontology engineering with deep learning. Semantic Web **15**(5), 1991–2004 (2024)
17. He, Y., Yuan, M., Chen, J., Horrocks, I.: Language models as hierarchy encoders. In: Globersons, A., et al. (eds.) Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, 10 - 15 December 2024 (2024). http://papers.nips.cc/paper_files/paper/2024/hash/1a970a3e62ac31c76ec3cea3a9f68fdf-Abstract-Conference.html
18. Jackermeier, M., Chen, J., Horrocks, I.: Dual box embeddings for the description logic el^{++}. In: Chua, T., Ngo, C., Kumar, R., Lauw, H.W., Lee, R.K. (eds.) Proceedings of the ACM on Web Conference 2024, WWW 2024, Singapore, 13-17 May 2024, pp. 2250–2258. ACM (2024).https://doi.org/10.1145/3589334.3645648
19. Kulmanov, M., Liu-Wei, W., Yan, Y., Hoehndorf, R.: EL embeddings: geometric construction of models for the description logic EL++. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, pp. 6103–6109. International Joint Conferences on Artificial Intelligence Organizatiohttps://doi.org/10.24963/ijcai.2019/845, https://www.ijcai.org/proceedings/2019/845
20. Kulmanov, M., Smaili, F.Z., Gao, X., Hoehndorf, R.: Semantic similarity and machine learning with ontologies. Briefings Bioinform. **22**(4), bbaa199 (2021)
21. Lee, J.: Introduction to Smooth Manifolds. Graduate Texts in Mathematics. Springer Science & Business Media (2013)
22. Lee, J.M.: Riemannian manifolds: an introduction to curvature, vol. 176. Springer Science & Business Media (2006)
23. Li, N., Bailleux, T., Bouraoui, Z., Schockaert, S.: Ontology completion with natural language inference and concept embeddings: an analysis. CoRR abs/arXiv: 2403.17216 (2024). https://doi.org/10.48550/ARXIV.2403.17216
24. Mondal, S., Bhatia, S., Mutharaju, R.: Emel++: embeddings for EL++ description logic. In: Martin, A., Hinkelmann, K., Fill, H., Gerber, A., Lenat, D., Stolle, R., van Harmelen, F. (eds.) Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021), Stanford University, Palo Alto, California, USA, March 22-24, 2021. CEUR Workshop Proceedings, vol. 2846. CEUR-WS.org (2021). https://ceur-ws.org/Vol-2846/paper19.pdf
25. Mungall, C.J., Torniai, C., Gkoutos, G.V., Lewis, S.E., Haendel, M.A.: Uberon, an integrative multi-species anatomy ontology. Genome Biol. **13**, 1–20 (2012)
26. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA. pp. 6338–6347 (2017). https://proceedings.neurips.cc/paper/2017/hash/59dfa2df42d9e3d41f5b02bfc32229dd-Abstract.html
27. Peng, X., Tang, Z., Kulmanov, M., Niu, K., Hoehndorf, R.: Description logic EL++ embeddings with intersectional closure. http://arxiv.org/abs/2202.14018
28. Rector, A.L., Rogers, J.E., Pole, P.: The galen high level ontology. In: Medical Informatics Europe 1996, pp. 174–178. IOS Press (1996)
29. Smaili, F.Z., Gao, X., Hoehndorf, R.: Opa2vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. Bioinform. **35**(12), 2133–2140 (2019)
30. Tang, Z., Hinnerichs, T., Peng, X., Zhang, X., Hoehndorf, R.: Falcon: faithful neural semantic entailment over alc ontologies. arXiv preprint arXiv:2208.07628 (2022)

31. Vaswani, A., et al.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pp. 5998–6008 (2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
32. Wang, X., et al.: KEPLER: a unified model for knowledge embedding and pretrained language representation. Trans. Assoc. Comput. Linguistics **9**, 176–194 (2021). https://doi.org/10.1162/TACL_A_00360
33. Xiong, B., Potyka, N., Tran, T.K., Nayyeri, M., Staab, S.: Faithiful embeddings for EL++ knowledge bases. https://arxiv.org/abs/2201.09919v2
34. Yang, H., Chen, J.: Regd: hierarchical embeddings via distances over geometric regions. arXiv preprint arXiv:2501.17518 (2025)
35. Yang, H., Chen, J., Sattler, U.: Transbox: EL^{++}-closed ontology embedding. In: The Web Conference (2025)
36. Yao, L., Mao, C., Luo, Y.: KG-BERT: BERT for knowledge graph completion. CoRR abs/ arXiv: 1909.03193 (2019). http://arxiv.org/abs/1909.03193
37. Zhapa-Camacho, F., Hoehndorf, R.: Cate: Embedding alc ontologies using category-theoretical semantics (2023)
38. Zhapa-Camacho, F., Kulmanov, M., Hoehndorf, R.: mowl: python library for machine learning with biomedical ontologies. Bioinformatics **39**(1), btac811 (2023)