Optimizing Class Subsumption through Controlled Dynamics of n-Balls in Vector Space

Aniket Mitra[1][0009−0007−6343−3830] and Vinu E. Venugopal[2][0000−0003−4429−9932]

International Institute of Information Technology, Bangalore, India
{aniket.mitra,vinu.ev}@iiitb.ac.in

Abstract. Representing entities from an ontology as geometric shapes (such as balls, boxes, etc.) in a low-dimensional vector space, known as *Region-based Geometric Knowledge Graph Embedding* or RKGE, has demonstrated the ability to outperform traditional knowledge graph embedding methods in reasoning tasks while preserving the structural properties and syntactic characteristics of ontological axioms. In this study, we introduce a novel approach to enhance the subsumption capability of geometric embeddings based on *n-balls*. Additionally, we propose techniques to enhance the quality of such embeddings by extracting meta-information from the information-rich lexicons or annotations within the domain ontology.

Keywords: $\mathcal{EL}^{++}$ DL · n-Ball Embedding · Knowledge Graph Embeddings

1 Introduction

Model theoretic languages like Description Logic (DL) are used to represent the semantics of OWL ontology axioms. $\mathcal{EL}$, a sub-language of DL, is widely used to represent large biomedical ontologies such as GO and SNOMED due to its fast reasoning (tractable) property and support of major symbolic logic constructs including concept intersection, existential relations between concepts, etc. Notably, the *general* TBox axioms of $\mathcal{EL}^{++}$ (an extension of $\mathcal{EL}$) can be reduced to one of the below normal forms (NF 1 to 4) in linear time maintaining resultant normalized TBox size linear to the original TBox thereby still guaranteeing tractable reasoning property [1].

– **NF 1-4 (Concept axioms):** $C \sqsubseteq D$, $C \sqcap D \sqsubseteq E$, $\exists R.C \sqsubseteq D$, $C \sqsubseteq \exists R.D$

The bottom concept axioms can be represented by assigning $\bot$ to the right-hand side of NF1-3. Role inclusion axioms can be normalized in linear time and represented as below. We named them as NF 5-7 for the ease of addressing.

– **NF 5 (Bottom concept axioms):** $C \sqcap D \sqsubseteq \bot$, $\exists R.C \sqsubseteq \bot$, $C \sqsubseteq \bot$

– **NF 6-7 (Role inclusion axioms):** $R \sqsubseteq S$, $R_1 \circ R_2 \sqsubseteq S$

where $\{C, D, E, \bot\} \in N_c$ & $\{R, R_1, R_2, S\} \in N_r$. Here N_c and N_r denote the set of classes and roles in the ontology respectively.

In EmEL^{++} model [5] (a.k.a. *n-ball* approach), an extension of *Region-based Geometric Knowledge Graph Embedding* (RKGE) model called ELEm model [4],

the authors attempted geometric construction of each concept and role as *balls* and translation vectors respectively in vector space $\mathbb{R}^n$. They formulated specific loss functions preserving the semantic meaning of each of these $\mathcal{EL}^{++}$ NFs and optimized the balls via a Machine Learning (ML) model trained on these loss functions. Proper training will assign close vector representations to subsumption balls and bigger radius to the super-concept ball such that the sub-concept ball is totally engulfed by it's super-concept ball. Later, [6] introduced *box-shape* to represent concepts since balls do not satisfy intersectional closure property. An alternative technique for embedding ontological data utilizes the graph-walk approach, albeit it ignore the structural nuances and characteristics inherent in the underlying ontology. In this method, each concept is represented as a node, while the relationships are illustrated as edges within the graph. Multiple rounds of randomized walks are executed on this graph structure to generate embeddings. Among these methodologies, OWL2Vec* [2] stands out by harnessing the substantial meta-information inherent in ontologies, including labels, synonyms, definitions, and more. By integrating this information into the graph structure, OWL2Vec* achieves a more comprehensive representation of the ontology, thereby enhancing its reasoning capabilities.

The ball method conducts reasoning operations in linear time utilizing a straightforward ML model with minimal hyper-parameters, as opposed to complex graph-walk models that entail multiple path explorations, extensive pre-trained language models, and numerous parameters. Nevertheless, the accuracy of RKGE is significantly contingent upon the design of its loss functions. However, the meta-information present in the ontology is largely ignored while fine-tuning the model. We hypothesize that since a sub-ball is nothing but the more specific version of it's super-ball, they must have common metadata terms that can be utilized to push the sub-balls properly inside their correct super-balls by tailoring accurate loss functions as depicted in Figure 2.

2 Proposed Approach: Meta-n-ball Model

Figure 1 outlines the general workflow of our proposed approach, *Meta-n-ball Model*[1]. We gather meta-information from the ontology and feed it through a pre-trained NLP model to extract biomedical terms. These terms are then initialized as n-dimensional (n-D) vectors in the metadata vector space, also known as *meta-space*. Simultaneously, concepts are initialized in the Concept Vector Space, or *con-space*, as n-D balls. Our meta-n-ball module subsequently refines these balls to generate the final embeddings.

In order to maintain the distinctiveness of the explicit concepts from both the ontology and the extracted meta-information, we have chosen to represent them in two separate vector spaces as mentioned above, drawing inspiration from [3]. The combined loss function for the con-space and the meta-space vectors contains three components as shown in Figure 1. The loss functions of the seven

[1] https://github.com/bda-lab/meta-n-ball

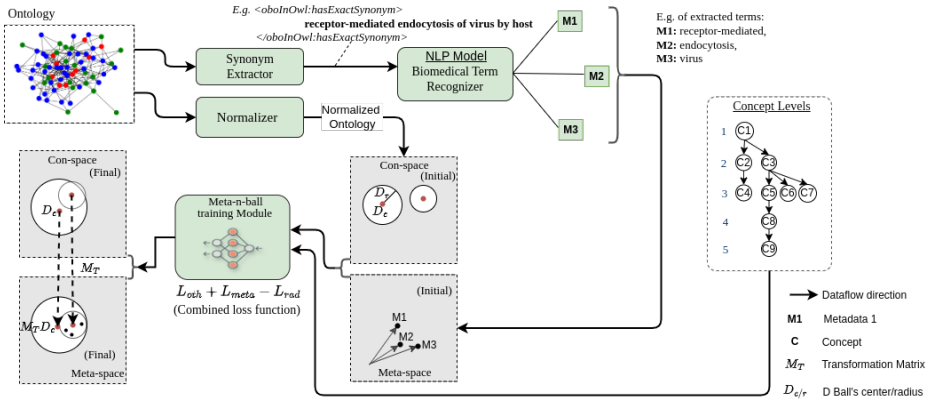

Fig. 1. Dataflow showing the generation of Meta-n-ball from domain ontology

NFs, as denoted collectively by L_{oth}, are designed for training the con-space entities based on [5]. Additionally, L_{rad} incorporates concept-level information. To enhance the quality of the ball's radius, the shortest distance to a specific concept from the root concept node in the concept hierarchy (a DAG) also called level is taken into consideration. This approach emphasizes the loss function more prominently for smaller balls. In Eq. 1, Le_i (0 if unavailable) denotes level of a concept with radius R_i and γ denotes margin loss parameter.

$$L_{rad} = \begin{cases} [R_i]_- + \gamma & \text{if } Le_i = 0 \\ \left[\sqrt{Le_i} * R_i\right]_- & \text{if } Le_i >= 1 \end{cases} \tag{1}$$

The loss function L_{meta}, representing the third type of loss function, is dedicated to learning meta-information. In Eq. 2, we use transformation matrix M_T which maps the concepts to meta-space and ensures that the metadata m resides inside it's correct concept C (center C_c & radius R_c) in meta-space.

$$L_{meta} = \left[\|M_T C_c - m\| - R_c - \gamma\right]_+ + \left|\|C_c\| - 1\right| + \left|\|m\| - 1\right| \tag{2}$$

In the equations (1) and (2), $[x]_+$, $[x]_-$ symbolizes $max(0, x)$ and $min(0, x)$ respectively. During training, the concepts, relations, metadata embeddings and M_T are optimized at every iteration in mini-batches based on the aggregation of these loss functions to reach their final values.

3 Results & Conclusion

Evaluation Metrics. The existing evaluation metrics primarily focus on calculating the distance between the centers of subsumption balls, often overlooking the quality aspects of the generated balls. This includes whether the radii are positive, whether the super-ball has a greater radius, the quality of subsumption (either *total* or *partial*, where total is preferred more), etc. Keeping these in mind

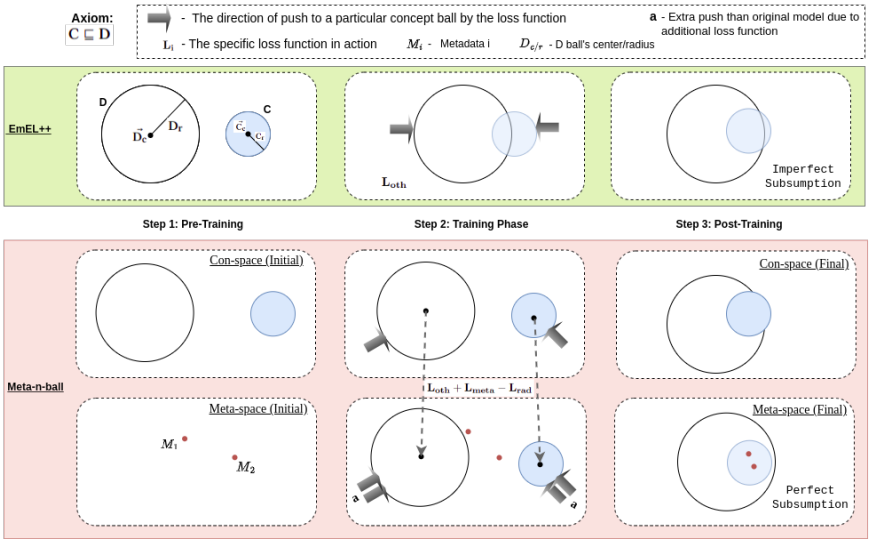

Fig. 2. Additional Loss Functions Improving Subsumption Quality

we design the following evaluation criteria: (1) *Valid Radius Proportion (VRP)* − The proportion of test cases where both radii are positive and the radius of the super-concept ball is larger than that of sub-concept. Naturally, the higher this proportion, the better the model is. (2) *Overall Distance between Centers (ODC)* − To assess whether the distance between the centers of sub and super balls has decreased across all test cases in our new model, we conducted a one-tailed t-test to determine which model exhibits greater distance values. We calculate the distance between the balls in meta-space in the case of the meta-n-ball model. The reporting format is as follows: if there is a significant reduction (p-val$<$0.05) in distance for the meta-n-ball model than EmEL^{++}, it is denoted as *(T-statistic, emel>meta-nball)*, and vice versa. The subsumed ball pairs are expected to have lesser distance between them for the better model. (3) *Perfect Overlapping Proportion (TOP)* − The proportion of valid VRP test cases where the sub-concept ball is completely inside it's super-concept ball in meta-space. Higher proportion is expected for better model. Here the term "valid" denotes all the test cases where both the sub and super balls have radius greater than 0 and super-ball's radius is greater than it's sub-ball. The format for reporting VRP and TOP are: *<proportion (count of proportion)>* Eg: *0.25 (500)*

Table 1. Comparing n-ball (EmEL++) and Meta-n-ball approaches.

Dataset	Evaluation Metrics	EmEL++	Meta-n-ball	EmEL++	Meta-n-ball	EmEL++	Meta-n-ball
Test Splits		Split1		Split2		Split3	
GO	VRP	0.707 (8391)	**0.742 (8813)**	0.596 (7066)	**0.640 (7596)**	0.587 (6971)	**0.628 (7456)**
	ODC(T-test)	**14.88, emel>meta-nball**		**69.19, emel>meta-nball**		**64.09, emel>meta-nball**	
	TOP	0.241 (2862)	**0.424 (5029)**	0.242 (2872)	**0.385 (4573)**	0.234 (2782)	**0.344 (4088)**
HPO	VRP	0.621 (6584)	**0.683 (7239)**	0.381 (3994)	**0.430 (4502)**	0.402 (4243)	**0.442 (4661)**
	ODC(T-test)	3.156, meta-nball>emel		**14.62, emel>meta-nball**		**5.71, emel>meta-nball**	
	TOP	0.203 (2153)	**0.238 (2521)**	0.091 (956)	**0.121 (1267)**	0.089 (938)	**0.140 (1468)**

Experiments & Results. We utilized the Python package *Scispacy* with the *en_core_sci_md* NLP model, pretrained on $50K$ relevant biomedical entities. Metadata was extracted from synonym data (*hasExactSynonym*, *hasRelatedSynonym*, *hasBroadSynonym*, *hasNarrowSynonym*) found in the Gene Ontology[2] (GO) and Human Phenotype Ontology[3] (HPO). We set hyper-parameters for both models as follows: $n = 100$, $\gamma = -0.1$, *epoch*$= 1000$ where n denotes the number of dimensions, γ denotes the margin loss parameter and *epoch* is the number of iterations the algorithms will run for. Concept levels were extracted from the respective *.obo* files available on official websites[2,3] using the Python *goatools* package. Our model was tested on three separate valid-test splits, as shown in Table 1, with each test sample containing the true subsumption relation in the format (<sub-ball> <super-ball>). Our meta-n-ball approach consistently enhances the quality of ball radii and rectifies numerous imperfect subsumptions, as evidenced by improved VRP and TOP values across all test splits in Table 1. ODC performance is also commendable across almost all test splits indicating meta-n-ball model has successfully reduced the distance between the centers of subsumption ball pairs. But, upon closer examination of extracted biomedical terms, it is apparent that some metadata are overly generic (e.g., "activity", "positive") or near-duplicates (e.g., "ureteric reflux", "ureteral reflux"). Hence, implementing filtering criteria to include only relevant terms is essential.

Conclusions. The experimental studies substantiate our hypothesis that the integration of metadata alongside meticulously designed loss functions can effectively enhance subsumption quality, thereby stimulating further exploration. Filtering out imperfect metadata and exploring additional meta-information, such as definitions and labels, while ensuring performance, presents an intriguing research challenge. This work suggests promising future research directions for advancing knowledge representation and reasoning in the RKGE framework.

References

1. Baader, F., Brandt, S., Lutz, C.: Pushing the EL envelope. In: Proceedings of the IJCAI. pp. 364–369. Professional Book Center (2005)
2. 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)
3. Hao, J., Chen, M., Yu, W., Sun, Y., Wang, W.: Universal representation learning of knowledge bases by jointly embedding instances and ontological concepts. In: Proceedings of the 25th ACM SIGKDD. pp. 1709–1719. ACM (2019). https://doi.org/10.1145/3292500.3330838
4. Kulmanov, M., Liu-Wei, W., Yan, Y., Hoehndorf, R.: EL embeddings: Geometric construction of models for the description logic EL++. arXiv preprint arXiv:1902.10499 (2019)
5. Mondal, S., Bhatia, S., Mutharaju, R.: EmEL++: Embeddings for EL++ description logic. In: Proceedings of the AAAI-MAKE. vol. 2846. CEUR-WS.org (2021)
6. Peng, X., Tang, Z., Kulmanov, M., Niu, K., Hoehndorf, R.: Description logic EL++ embeddings with intersectional closure. arXiv preprint arXiv:2202.14018 (2022)

[2]https://geneontology.org/docs/download-ontology/
[3]https://hpo.jax.org/app/data/annotations