Towards Statistical Reasoning with Ontology Embeddings

Yuqicheng Zhu[1,2,*,†], Nico Potyka[3,†], Bo Xiong[1], Trung-Kien Tran[2], Mojtaba Nayyeri[1], Steffen Staab[1] and Evgeny Kharlamov[2]

[1]University of Stuttgart, Stuttgart, Germany
[2]Bosch Center for Artificial Intelligence, Renningen, Germany
[3]Cardiff University, Cardiff, UK

Abstract

The Description Logic (DL) $\mathcal{EL}$ is a lightweight DL that has a favorable trade-off between expressive power and reasoning complexity and has been widely used in many real-world applications. Statistical $\mathcal{EL}$ ($\mathcal{SEL}$) extends $\mathcal{EL}$ by allowing conditional probabilities over axioms. Unlike other probabilistic DLs, the probabilistic semantics of $\mathcal{SEL}$ is statistical, meaning that probabilities express proportions in a population rather than subjective beliefs. One major challenge is that reasoning in $\mathcal{SEL}$ is ExpTime-complete. To overcome this problem, we propose to use embeddings to perform approximate inference over $\mathcal{SEL}$ ontologies. This poster paper demonstrates the progress of the ongoing research, showcasing a demonstration through a simplified example, providing preliminary findings, and outlining the future work.

Keywords

Description logics, Uncertain reasoning, Ontology embeddings

1. Introduction

Description logics (DLs) [1] are logical languages used for representing ontological knowledge. Different DLs balance expressive power and reasoning complexity. One of the most prominent DLs is the Existential Language ($\mathcal{EL}$) [2], which supports conjunction and existential quantification. $\mathcal{EL}$ is sufficiently expressive for most ontologies that occur in practice and has polynomial reasoning complexity. Due to its appealing properties, $\mathcal{EL}$ has become one of the underlying formalisms of the standardized Web Ontology Language (OWL2 EL) [3].

In $\mathcal{EL}$ ontologies, knowledge is expressed by subsumption relationships like $\mathit{Politician} \sqsubseteq \mathit{Person}$ meaning that politicians are persons. However, there is usually uncertainty about our real-world knowledge. Statistical $\mathcal{EL}$ [4] ($\mathcal{SEL}$) is a statistical variant of $\mathcal{EL}$ that allows reasoning about statistics of a population. $\mathcal{SEL}$ ontologies are composed of probabilistic conditionals of the form $(D \mid C)[l, u]$, where $0 \leq l \leq u \leq 1$. For example, $(\mathit{Politician} \mid \mathit{Doctor})[0.1, 0.2]$ expresses that around 10-20% of persons with doctorate degree are politicians. Unfortunately, reasoning in $\mathcal{SEL}$ is ExpTime-complete [5] and, therefore, provably intractable.

ISWC 2023 Posters and Demos: 22nd International Semantic Web Conference, November 6–10, 2023, Athens, Greece
*Corresponding author.
†These authors contributed equally.
yuqicheng.zhu@ipvs.uni-stuttgart.de (Y. Zhu); PotykaN@cardiff.ac.uk (N. Potyka)

To overcome the practical limitations, we propose using embeddings to perform approximate reasoning in $\mathcal{SEL}$. The intuitive idea is to map concepts to boxes in a vector space. The statistical proportions in the ontology are maintained by guaranteeing similar proportions between the volume of the boxes. We can then reason about arbitrary concepts by computing new proportions between volumes in the vector space. Technically, we generalize the $\mathcal{EL}$ embedding BoxEL from [6] to $\mathcal{SEL}$. In practice, the ontology is typically not perfectly represented by the embedding. However, we assume that the approximation error is small whenever the embedding error is small. To evaluate this empirically, we derive a sound inference rule (Probabilistic Modus Ponens) for $\mathcal{SEL}$ and use it to evaluate the approximation quality of our approach empirically. Our experiments show that both the embedding and inference errors are typically very small.

2. Embedding and Approximating $\mathcal{SEL}$

The DL $\mathcal{EL}$ [7] describes individuals, concepts and their relationships using a set N_I of *individual names*, a set N_C of *concept names* and a set N_R of *role/relation names*. Roughly speaking, $\mathcal{EL}$ allows talking about concepts that are formed from atomic concepts by taking conjunctions (denoted by $\sqcap$) and existential quantification (denoted by $\exists$). Existential quantification in DLs assures the existences of a role successor. For example, $\exists \mathit{has}.\mathit{Child}$ refers to the set of objects that have (role) a child (concept). $\mathcal{EL}$ TBoxes are collections of subsumption relationships between concepts that are called general concept inclusions (GCIs). A GCI has the form $C \sqsubseteq D$, where C and D are $\mathcal{EL}$ concepts.

$\mathcal{SEL}$ is a probabilistic extension of $\mathcal{EL}$ that allows reasoning about statistical statements [4]. The basic syntactic elements are *(probabilistic) conditionals* $(D \mid C)[l, u]$, where C, D are $\mathcal{EL}$ concept descriptions and l, u are probabilities such that $l \leq u$. Intuitively, $(D \mid C)[l, u]$ expresses that the proportion of individuals in C that also belong to D is between l and u. If $l = u$, we simplify notation and just write $(D \mid C)[l]$. $\mathcal{SEL}$ generalizes $\mathcal{EL}$ in the sense that $(D \mid C)[1]$ is semantically equivalent to $C \sqsubseteq D$ [4]. To illustrate the additional expressiveness of $\mathcal{SEL}$, let us consider some statistical beliefs about food.

$(\mathit{SpicyPizza} \mid \mathit{Food})[0.4]$	$(\mathit{CheesyPizza} \mid \mathit{Food})[0.45]$
$(\mathit{IceCream} \mid \mathit{Food})[0.1]$	$(\mathit{Cake} \mid \mathit{Food})[0.1]$
$(\mathit{SpicyPizza} \mid \mathit{IceCream})[0]$	$(\mathit{Food} \mid \mathit{IceCream})[1]$
$(\mathit{SpicyPizza} \sqcap \mathit{CheesyPizza} \mid \mathit{MeatyPizza})[0.4]$	$(\mathit{SpicyPizza} \sqcap \mathit{MeatyPizza} \mid \mathit{CheesyPizza})[0.35]$
$(\mathit{eatWith.IceCream} \mid \mathit{Food})[0.25]$	$(\mathit{CheesyPizza} \mid \mathit{eatWith.IceCream})[0.9]$

The first two rows state proportions of different food products (they do not have to be disjoint and so the probabilities do not need to sum up to 1). The third row represents deterministic knowledge, stating the disjointness of ice cream and spicy pizza, and the fact that ice cream is always classified as food. The last two rows show some more complex examples with conjunction and existential quantification. For instance, the conditional $(\mathit{eatWith.IceCream} \mid \mathit{Food})[0.25]$ expresses that 25% of food products are eaten with ice cream.

Given an arbitrary $\mathcal{SEL}$ conditional $(D \mid C)[l, u]$, we first replace it with the conditional $(B \mid A)[l, u]$, where A, B are new concept names corresponding to C and D. To guarantee equivalence

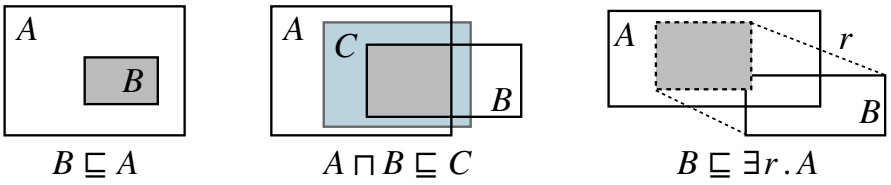

Figure 1: The geometric interpretation of logic statements in TBox expressed by $\mathscr{EL}++$[6]

$C \equiv A$ and $D \equiv B$, we add four $\mathscr{EL}$ GCIs $A \sqsubseteq C$, $C \sqsubseteq A$, $B \sqsubseteq D$, $D \sqsubseteq B$. To perform approximate inference on this knowledge bases, we embed GCIs using the geometric interpretation in BoxEL[6]. Concretely, concepts are modeled as *boxes* (i.e., axis-aligned hyperrectangles) and the relation as the *affine transformation* between boxes (see figure. 1). Furthermore, we embed the remaining atomic conditionals by additional loss terms that encourage that the ratio between the intersection of box A and B, and A, respects the bounds expressed by the conditional. Having computed the embedding, we can perform approximate inference by computing unknown proportions in the embedding space.

3. Preliminary Experimental Results

Figure 2 shows a 2-dimensional embedding of a superset of our food ontology. Notably, cake, ice cream and pizzas are disjoint (due to additional disjointness constraints) and the proportion of meaty pizzas that are both spicy and cheesy appears visually accurate. Intuitively, we can infer new knowledge by considering unknown proportions in the embedding space. We expect that the inference error is small whenever the embedding error was small. To evaluate this hypothesis empirically, we automatically extract statistical conditionals from YAGO3 [8] by computing statistics for a collection of template conditionals. Specifically, we assess the inference capabilities across datasets with diverse characteristics namely "Person," "Country," and "Hybrid" using the following inference rule for $\mathcal{SEL}$.

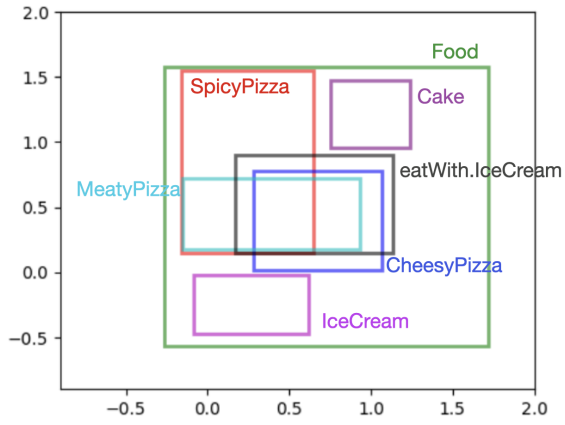

Figure 2: 2D BoxSEL embeddings of food example

Proposition 1 (Probabilistic Modus Ponens (PMP)). *If* $(C)[l_1, u_1]$ *and* $(D \mid C)[l_2, u_2]$, *then* $(D)[l_3, u_3]$, *where* $l_3 = l_1 \cdot l_2$ *and* $u_3 = \min\{1, u_1 \cdot u_2 + 1 - l_1\}$.

Preliminary experimental results are presented in table 1. The table is mostly in line with our assumption that if the embedding error (the loss of the logical terms in the embedding) is small, the inference error is small as well. In future work, we will try to make an analytic connection

Table 1
Evaluation of total embedding (EE) and inference (IE) error based on PMP and individual errors for the four template conditional types extracted from Yago (1 - atomic, 2 - conjunction, 3 and 4 existential).

	EE	IE	EE T1	IE T1	EE T2	IE T2	EE T3	IE T3	EE T4	IE T4
Person	0.049	0.012	0.029	0.009	0.096	0.026	0.002	0.001	0.039	0.008
Country	0.015	0.002	0.018	0.004	0.014	0.002	0.048	0.008	0.016	0.005
Hybrid	0.034	0.015	0.161	0.086	0.036	0.018	0.016	0.006	0.044	0.018

between the embedding error (that is known after computing the embedding) and the potential inference error (which is unknown).

4. Related Work

Our work builds up on knowledge graph (KG) embeddings that map entities and relations into a vector space to model the relationships between entities. Most KG embeddings encode factual/instance-level knowledge expressed by triples $\langle$*head entity*, *relation*, *tail entity*$\rangle$ but ignore the terminological/concept-level knowledge expressed by logical axioms. [9] proposed embedding $\mathscr{EL}$ concepts as n-balls and relations as *translations* between them. However, as balls are not closed under intersection, they cannot faithfully represent concept intersection. BoxEL [6] and ELBE [10] overcome this issue by embedding concepts as axis-parallel boxes. ELBE models relations as *translations* while BoxEL replaces *translations* by *affine transformations*. Box2EL [11] further considers one-to-many and many-to-many relations and embeds both concepts and roles as boxes. However, none of these methods is based on the probabilistic semantics that underlies $\mathcal{SEL}$ ontologies.

5. Discussion and Outlook

This poster paper presents our ongoing research focused on utilizing box embeddings for approximate inference over $\mathcal{SEL}$ ontologies. Our preliminary experiments show a small embedding and inference error and indicate that the known embedding error can be used to bound the unknown inference error. We are planning to utilize this in future work by reporting confidence intervals rather than point probabilities for queries. A related avenue for future investigation involves exploring the possibility of generating embeddings uniformly at random to get a better representation of the entailed probability interval (while our embedding always returns point probabilities, $\mathcal{SEL}$ knowledge bases typically entail interval probabilities). Additionally, incorporating region-based role embeddings, as proposed in [11], may help to reduce the embedding error and consequently the inference error further.

References

[1] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, P. F. Patel-Schneider (Eds.), The Description Logic Handbook: Theory, Implementation, and Applications, Cambridge

University Press, 2003.
[2] F. Baader, B. Morawska, Unification in the description logic el., in: RTA, Springer, 2009, pp. 350–364.
[3] B. C. Grau, I. Horrocks, B. Motik, B. Parsia, P. Patel-Schneider, U. Sattler, Web semantics: Science, services and agents on the world wide web, Web Semantics: Science, Services and Agents on the World Wide Web 6 (2008) 309–322.
[4] R. Peñaloza, N. Potyka, Towards statistical reasoning in description logics over finite domains, in: International Conference on Scalable Uncertainty Management (SUM), Springer, 2017, pp. 280–294.
[5] B. Bednarczyk, Statistical EL is exptime-complete, Information Processing Letters 169 (2021) 106113.
[6] B. Xiong, N. Potyka, T. Tran, M. Nayyeri, S. Staab, Faithful embeddings for $E\ell^{++}$ knowledge bases, in: U. Sattler, A. Hogan, C. M. Keet, V. Presutti, J. P. A. Almeida, H. Takeda, P. Monnin, G. Pirrò, C. d'Amato (Eds.), International Semantic Web Conference (ISWC), volume 13489 of *Lecture Notes in Computer Science*, Springer, 2022, pp. 22–38.
[7] F. Baader, Least common subsumers and most specific concepts in a description logic with existential restrictions and terminological cycles, in: International Joint Conference on Artificial Intelligence (IJCAI), Morgan Kaufmann, 2003, pp. 319–324.
[8] F. Mahdisoltani, J. Biega, F. Suchanek, Yago3: A knowledge base from multilingual wikipedias, in: 7th biennial conference on innovative data systems research, CIDR Conference, 2014.
[9] M. Kulmanov, W. Liu-Wei, Y. Yan, R. Hoehndorf, EL embeddings: Geometric construction of models for the description logic EL++, in: IJCAI, ijcai.org, 2019, pp. 6103–6109.
[10] X. Peng, Z. Tang, M. Kulmanov, K. Niu, R. Hoehndorf, Description logic EL++ embeddings with intersectional closure, CoRR abs/2202.14018 (2022).
[11] M. Jackermeier, J. Chen, I. Horrocks, Box^{2}el: Concept and role box embeddings for the description logic EL++, CoRR abs/2301.11118 (2023).