Improved Knowledge Graph Embedding using Background Taxonomic Information

Bahare Fatemi, Siamak Ravanbakhsh, and David Poole
The University of British Columbia
Vancouver, BC, V6T 1Z4
{bfatemi, siamakx, poole}@cs.ubc.ca

Abstract

Knowledge graphs are used to represent relational information in terms of triples. To enable learning about domains, embedding models, such as tensor factorization models, can be used to make predictions of new triples. Often there is background taxonomic information (in terms of subclasses and subproperties) that should also be taken into account. We show that existing fully expressive (a.k.a. universal) models cannot provably respect subclass and subproperty information. We show that minimal modifications to an existing knowledge graph completion method enables injection of taxonomic information. Moreover, we prove that our model is fully expressive, assuming a lower-bound on the size of the embeddings. Experimental results on public knowledge graphs show that despite its simplicity our approach is surprisingly effective.

The AI community has long noticed the importance of structure in data. While traditional machine learning techniques have been mostly focused on feature-based representations, the primary form of data in the subfield of Statistical Relational AI (StaRAI) (Getoor and Taskar, 2007; Raedt et al., 2016) is in the form of entities and relationships among them. Such entity-relationships are often in the form of (head, relationship, tail) triples, which can also be expressed in the form of a graph, with nodes as entities and labeled directed edges as relationships among entities. Predicting the existence, identity, and attributes of entities and their relationships are among the main goals of StaRAI.

Knowledge Graphs (KGs) are graph structured knowledge bases that store facts about the world. A large number of KGs have been created such as NELL (Carlson et al., 2010), FREEBASE (Bollacker et al., 2008), and Google Knowledge Vault (Dong et al., 2014). These KGs have applications in several fields including natural language processing, search, automatic question answering and recommendation systems. Since accessing and storing all the facts in the world is difficult, KGs are incomplete. The goal of *link prediction* for KGs – a.k.a. KG *completion* – is to predict the unknown links or relationships in a KG based on the existing ones. This often amounts to infer (the probability of) new triples from the existing triples.

A common approach to apply machine learning to symbolic data, such as text, graph and entity-relationships, is through embeddings. Word, sentence and paragraph embeddings (Mikolov et al., 2013; Pennington, Socher, and Manning, 2014), which vectorize words, sentences and paragraphs using context information, are widely used in a variety of natural language processing tasks from syntactic parsing to sentiment analysis. Graph embeddings (Hoff, Raftery, and Handcock, 2002; Grover and Leskovec, 2016; Perozzi, Al-Rfou, and Skiena, 2014) are used in social network analysis for link prediction and community detection.

In relational learning, embeddings for entities and relationships are used to generalize from existing data. These embeddings are often formulated in terms of tensor factorization (Nickel, Tresp, and Kriegel, 2012; Bordes et al., 2013; Trouillon et al., 2016; Kazemi and Poole, 2018c). Here, the embeddings are learned such that their interaction through (tensor-)products best predicts the (probability of the) existence of the observed triples; see (Nguyen, 2017; Wang et al., 2017) for details and discussion. Tensor factorization methods have been very successful, yet they rely on a large number of annotated triples to learn useful representations. There is often other information in ontologies which specifies the meaning of the symbols used in a knowledge base. One type of ontological information is represented in a hierarchical structure called a taxonomy. For example, a knowledge base might contain information that DJTrump, whose name is "Donald Trump" is a president, but may not contain information that he is a person, a mammal and an animal, because these are implied by taxonomic knowledge. Being told that mammals are chordates, lets us conclude that DJTrump is also a chordate, without needing to have triples specifying this about multiple mammals. We could also have information about subproperties, such as that being president is a subproperty of "managing", which in turn is a subproperty of "interacts with".

This paper is about combining taxonomic information in the form of subclass and subproperty (*e.g.*, managing implies interaction) into relational embedding models. We show that existing factorization models that are fully expressive cannot reflect such constraints for all legal entity embeddings. We propose a model that is provably fully expressive and can represent such taxonomic information, and evaluate its performance on real-world datasets.

Factorization and Embedding

Let $\mathcal{E}$ represent the set of entities and $\mathcal{R}$ represent the set of relations. Let $\mathcal{W}$ be a set of *triples* (h, r, t) that are true in the world, where $h, t \in \mathcal{E}$ are *head* and *tail*, and $r \in \mathcal{R}$ is the *relation* in the triple. We use $\mathcal{W}^{\mathsf{c}}$ to represent the triples that are false – *i.e.*, $\mathcal{W}^{\mathsf{c}} \doteq \{(h, r, t) \in \mathcal{E} \times \mathcal{R} \times \mathcal{E} \mid (h, r, t) \notin \mathcal{W}\}$. An example of a triple in $\mathcal{W}$ can be (Paris, CapitalCityOfCountry, France) and an example of a triple in $\mathcal{W}^{\mathsf{c}}$ can be (Paris, CapitalCityOfCountry, Germany). A KG $\mathcal{K} \subseteq \mathcal{W}$ is a subset of all the facts. The problem of the KG completion is to infer $\mathcal{W}$ from its subset KG. There exists a variety of methods for KG completion. Here, we consider embedding methods and in particular using tensor-factorization. For a broader review of the existing KG completion that can use background information see Related Work.

Embeddings: An embedding is a function from an entity or a relation to a vector (or sometimes higher order tensors) over a field. We use bold lower-case for vectors – that is $\mathbf{s} \in \mathbb{R}^k$ is an embedding of an entity and $\mathbf{r} \in \mathbb{R}^l$ is an embedding of a relation.

Taxonomies: It is common to have structure over the symbols used in the triples, see (*e.g.*, Shoham, 2016). The Ontology Web Language (OWL) (Hitzler et al., 2012) defines (among many other meta-relations) subproperties and subclasses, where p_1 is a subproperty of p_2 if $\forall x, y : (x, p_1, y) \rightarrow (x, p_2, y)$, that is whenever p_1 is true, p_2 is also true. Classes can be defined either as a set with a class assertion (often called "type") between an entity and a class, *e.g.*, saying x is in class C using $(x, type, C)$ or in terms of the characteristic function of the class, a function that is true of element of the class. If c is the characteristic function of class C, then x is in class c is written $(x, c, true)$. For representations that treat entities and properties symmetrically, the two ways to define classes are essentially the same. C_1 is a subclass of C_2 if every entity in class C_1 is in class C_2, that is, $\forall x : (x, type, C_1) \rightarrow (x, type, C_2)$ or $\forall x : (x, c_1, true) \rightarrow (x, c_2, true)$. If we treat $true$ as an entity, then subclass can be seen as a special case of subproperty. For the rest of the paper we will refer to subsumption in terms of subproperty (and so also of subclass). A non-trivial subsumption is one which is not symmetric; p_1 is a subproperty of p_2 and there is some relations that is true of p_1 that is not true of p_2. We want the subsumption to be over all possible entities; those entities that have a legal embedding according to the representation used, not just those we know exist. Let $\mathcal{E}^*$ be the set of all possible entities with a legal embedding according to the representation used.

Tensor factorization: For KG completion a tensor factorization defines a function $\mu : \mathbb{R}^k \times \mathbb{R}^l \times \mathbb{R}^k \rightarrow [0, 1]$ that takes the embeddings $\mathbf{h}$, $\mathbf{r}$ and $\mathbf{t}$ of a triple (h, r, t) as input, and generates a prediction, *e.g.*, a probability, of the triple being true $(h, r, t) \in \mathcal{W}$. In particular, μ is often a non-linearity applied to a multi-linear function of $\mathbf{h}, \mathbf{r}, \mathbf{t}$. The family of methods that we study uses the following multi-linear form: Let $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ be vectors of length k. Define $\langle \mathbf{x}, \mathbf{y}, \mathbf{z} \rangle$ to be the sum of their element-wise product, namely

$$\langle \mathbf{x}, \mathbf{y}, \mathbf{z} \rangle \doteq \sum_{\ell=1}^{k} \mathbf{x}_\ell \mathbf{y}_\ell \mathbf{z}_\ell \tag{1}$$

where $\mathbf{x}_\ell$ is the ℓ-th element of vector $\mathbf{x}$.

Here, we are interested in creating a tensor-factorization method that is fully expressive and can incorporate background information in the form of taxonomy. A model is *fully expressive* if given any assignment of truth values to all triples, there exists an assignment of values to the embeddings of the entities and relations that accurately separates the triples belonging to $\mathcal{W}$ and $\mathcal{W}^{\mathsf{c}}$ using μ.

ComplEx

ComplEx (Trouillon et al., 2016) defines the reconstruction function μ, such that the embedding of each entity and each relation is a vector of complex numbers. Let $\mathrm{Re}(\mathbf{x})$ and $\mathrm{Im}(\mathbf{x})$ denote the real and imaginary part of a complex vector $\mathbf{x}$. In ComplEx, the probability of any triple (h, r, t) is

$$\mu(\mathbf{h}, \mathbf{r}, \mathbf{t}) \doteq \sigma(\mathrm{Re}(\langle \mathbf{h}, \mathbf{r}, \overline{\mathbf{t}} \rangle)) \tag{2}$$

where $\sigma : \mathbb{R} \rightarrow [0, 1]$ is the sigmoid or logistic function, and $\overline{\mathbf{a} + i\mathbf{b}} \doteq \mathbf{a} - i\mathbf{b}$ (where $i = \sqrt{-1}$) is the element-wise conjugate of the complex vector $\mathbf{a} + i\mathbf{b}$. Note that, if the tail did not use the conjugate, the head and tail would be treated symmetrically and it could only represent symmetric relations; *e.g.*, see DistMult in Yang et al. (2014).

Trouillon et al. (2017) prove that ComplEx is fully expressive. In particular, they prove that any assignment of ground truth can be modeled by ComplEx embeddings of length $|\mathcal{E}||\mathcal{R}|$. The following theorem shows that we cannot use ComplEx to enforce our prior knowledge about taxonomies.

Theorem 1 *ComplEx cannot enforce non-trivial subsumption.*

Proof *Assume a non-trivial subsumption so that* $\forall h, t \in \mathcal{E}^* : (h, r, t) \rightarrow (h, s, t)$*, and so* $\mu(\mathbf{h}, \mathbf{s}, \mathbf{t}) \geq \mu(\mathbf{h}, \mathbf{r}, \mathbf{t})$*, and there are entities* $a, b \in \mathcal{E}^*$ *such that* $\mu(\mathbf{a}, \mathbf{s}, \mathbf{b}) > \mu(\mathbf{a}, \mathbf{r}, \mathbf{b})$*. Let* a' *be an entity such that* $\mathbf{a}' = -\mathbf{a}$*. Then* $\mu(\mathbf{a}', \mathbf{s}, \mathbf{b}) = 1 - \mu(\mathbf{a}, \mathbf{s}, \mathbf{b})$ *and* $\mu(\mathbf{a}', \mathbf{r}, \mathbf{b}) = 1 - \mu(\mathbf{a}, \mathbf{r}, \mathbf{b})$*, so* $\mu(\mathbf{a}', \mathbf{s}, \mathbf{b}) < \mu(\mathbf{a}', \mathbf{r}, \mathbf{b})$*, a contradiction to the subsumption we assumed.* ■

Recently, Ding et al. (2018) proposed a method which they call ComplEx-NNE+AER to incorporate a weaker notion of subsumption in ComplEx. For a subsumption $\forall h, t \in \mathcal{E}^* : (h, r, t) \rightarrow (h, s, t)$, they suggest adding soft constraints to the loss function to encourage $\mathrm{Re}(\mathbf{r}) \leq \mathrm{Re}(\mathbf{s})$ and $\mathrm{Im}(\mathbf{r}) = \mathrm{Im}(\mathbf{s})$. When the constraints are satisfied, ComplEx-NNE+AER ensures $\forall h, t \in \mathcal{E} : \mu(\mathbf{h}, \mathbf{r}, \mathbf{t}) \leq \mu(\mathbf{h}, \mathbf{s}, \mathbf{t})$. This is a weaker notion than the definition in the Factorization and Embedding section which requires $\forall h, t \in \mathcal{E} : \mu(\mathbf{h}, \mathbf{r}, \mathbf{t}) \leq \mu(\mathbf{h}, \mathbf{s}, \mathbf{t})$ (that is, $\mathcal{E}^*$ is replaced with $\mathcal{E}$).

Theorem 2 *ComplEx-NNE+AER cannot satisfy its constraints and be fully expressive if symmetry constraints are allowed.*

Table 1: Results for the choice of non-linearity in producing non-negative embeddings.

	WN18					FB15k				
	MRR		Hit@			MRR		Hit@		
Function f	Filter	Raw	1	3	10	Filter	Raw	1	3	10
SimplE$^+$-Exponential	0.866	0.547	0.829	0.925	0.897	0.575	0.248	0.468	0.640	0.773
SimplE$^+$-Logistic	0.854	0.542	0.836	0.863	0.885	0.425	0.228	0.294	0.491	0.694
SimplE$^+$-ReLU	0.937	0.575	0.936	0.938	0.939	0.725	0.240	0.658	0.770	0.841

Proof *In ComplEx a relation r is symmetric for all possible entities if and only if $\mathrm{Im}(\mathbf{r}) = 0$ (Trouillon et al., 2016, Section 3). In order to satisfy constraints for $\forall h,t \in \mathcal{E} : (h,r,t) \rightarrow (h,s,t)$, Ding et al. (2018) assign $\mathrm{Im}(\mathbf{r}) = \mathrm{Im}(\mathbf{s})$. Therefore, if relation r is symmetric, it enforces relation s to be symmetric too which is not generally true. As a counter example, r might be the married_to relation, which is symmetric (so the $\mathrm{Im}(\boldsymbol{married\_to}) = 0$), but s is the knows relation, and $\forall h,t \in \mathcal{E} : (h,married\_to,t) \rightarrow (h,knows,t)$ is true in real-world, but setting the $\mathrm{Im}(\boldsymbol{knows}) = \mathrm{Im}(\boldsymbol{married\_to})$ will imply knows is symmetric, which is not true (as many people know celebrities but celebrities do not know many people).* ■

SimplE

SimplE (Kazemi and Poole, 2018c) achieves state-of-the-art in KG completion by considering two embeddings for each relation: one for the relation $r \in \mathcal{R}$ itself and one for its inverse. We use $\mathbf{r}^+ \in \mathbb{R}^k$ to denote the "forward" embedding of r and $\mathbf{r}^- \in \mathbb{R}^k$ to denote the embedding of its inverse. The embedding $\mathbf{r} = [\mathbf{r}^+, \mathbf{r}^-]$ for a relation is a concatenation of these two parts. Similarly, the embedding for each entity $e \in \mathcal{E}$ has two parts: its embedding as a head $\mathbf{e}^+$ and as a tail $\mathbf{e}^-$ – that is $\mathbf{e} = [\mathbf{e}^+, \mathbf{e}^-]$. Using this notation, SimplE calculates the probability of $(h,r,t) \in \mathcal{W}$ for each triple in both forward and backward directions using

$$\mu(\mathbf{h},\mathbf{r},\mathbf{t}) \doteq \sigma\left(\frac{1}{2}(\langle \mathbf{h}^+, \mathbf{r}^+, \mathbf{t}^+\rangle + \langle \mathbf{t}^-, \mathbf{r}^-, \mathbf{h}^-\rangle)\right). \quad (3)$$

Kazemi and Poole (2018c) prove SimplE is fully expressive and provide a bound on the size of the embedding vectors: For any truth assignment $\mathcal{W} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, there exists a SimplE model with embedding vectors of size $\min(|\mathcal{E}\,||\mathcal{R}\,|, |\mathcal{W}| + 1)$ that represent the assignment. The following theorem shows the limitation of SimplE when it comes to enforcing subsumption.

Theorem 3 *SimplE cannot enforce non-trivial subsumptions.*

Proof *Consider $\forall h,t \in \mathcal{E}^* : (h,r,t) \rightarrow (h,s,t)$ as a non-trivial subsumption. So we have $\mu(\mathbf{h},\mathbf{s},\mathbf{t}) \geq \mu(\mathbf{h},\mathbf{r},\mathbf{t})$, and there are entities $a,b \in \mathcal{E}^*$ such that $\mu(\mathbf{a},\mathbf{s},\mathbf{b}) > \mu(\mathbf{a},\mathbf{r},\mathbf{b})$. Let a' be an entity such that $\mathbf{a}' = -\mathbf{a}$. Then $\mu(\mathbf{a}',\mathbf{s},\mathbf{b}) = 1 - \mu(\mathbf{a},\mathbf{s},\mathbf{b})$ and $\mu(\mathbf{a}',\mathbf{r},\mathbf{b}) = 1 - \mu(\mathbf{a},\mathbf{r},\mathbf{b})$, so $\mu(\mathbf{a}',\mathbf{s},\mathbf{b}) < \mu(\mathbf{a}',\mathbf{r},\mathbf{b}))$ a contradiction to the subsumption we assumed.* ■

Table 2: Statistics on the datasets.

Dataset	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{R}\rvert$	#train	#valid	#test
WN18	40,943	18	141,442	5,000	5,000
FB15k	14,951	1,345	483,142	50,000	59,071
Sport	1039	5	1312	-	307
Location	445	5	384	-	100

Neural network models

The neural network models (Socher et al., 2013; Dong et al., 2014; Santoro et al., 2017) are very flexible, and so without explicit mechanisms to enforce subsumption, they cannot be guaranteed to obey any subsumption knowledge.

Proposed Variation: SimplE$^+$

In this section we propose a slight modification on SimplE so that the resulting method can enforce subsumption. The modification is restricting entity embeddings to be non-negative – that is $\mathbf{e}^+, \mathbf{e}^- \geq 0 \ \forall e \in \mathcal{E}$, where the inequality is element-wise. Next we show that the resulting model is fully expressive and is able to enforce subsumption.

Theorem 4 (Expressivity) *For any truth assignment over entities $\mathcal{E}$ and relations $\mathcal{R}$ containing $|\mathcal{W}|$ true facts, there exists a SimplE$^+$ model with embeddings vectors of size $\min(|\mathcal{E}\,||\mathcal{R}\,|+1, |\mathcal{W}|+1)$ that represent the assignment.*

Proof *Assume r_i is the i-th relation in $\mathcal{R}$ and e_j is the j-th entity in $\mathcal{E}$. For a vector a we define $(a)_i$ as the i-th element of a. We define $(\mathbf{r}_i^+)_n = 1$ if $n \operatorname{div} |\mathcal{E}| = i$ except the last element $(\mathbf{r}_i^+)_{|\mathcal{E}||\mathcal{R}|} = -1$, and for each entity s_i we define $(\mathbf{s}_j^+)_n = 1$ if $n \bmod |\mathcal{E}| = j$ or $n = |\mathcal{E}||\mathcal{R}|$ and 0 otherwise. In this setting, for each r_i and e_j product of $\mathbf{r}_i$ and $\mathbf{s}_i$ is 0 everywhere except for the element at $(i*|\mathcal{E}|+j)$ and the last element in the embeddings. In order for the triple (e_j, r_i, e_k) to hold, we define $(\mathbf{s}_k^-)$ to be a vector where all elements are 0 except the $(i*|\mathcal{E}|+j)$-th element which is 2. This proves that SimplE$^+$ is fully expressive with the bound of $|\mathcal{E}\,||\mathcal{R}\,|+1$ for size of the embeddings.*

We use induction to prove the bound $|\mathcal{W}|+1$. Let $|\mathcal{W}| = 0$ (base of induction). We can have embedding vectors of size 1 for each entity and relation, setting the value for entities to 1 and to relations to -1. Then $\langle \mathbf{h}^+, \mathbf{r}^+, \mathbf{t}^+\rangle + \langle \mathbf{t}^-, \mathbf{r}^-, \mathbf{h}^-\rangle$ is negative for every entities h and t and relation r. So there exist an assignment of size 1 that represent this ground truth.

Let's assume for any ground truth where $|\mathcal{W}| = n-1$, there exists an assignment of values to embedding vectors of size

Table 3: Relations and Rules in Sport and Location datasets.

	Relations	Subsumptions
Sport	AthleteLedSportsTeam	$(x, AtheleLedSportsTeam, y) \rightarrow (x, AthletePlaysForTeam, y)$
	AthletePlaysForTeam	$(x, AthletePlaysForTeam, y) \rightarrow (x, PersonBelongsToOrganization, y)$
	CoachesTeam	$(x, CoachesTeam, y) \rightarrow (x, PersonBelongsToOrganization, y)$
	OrganizationHiredPerson	$(x, OrganizationHiredPerson, y) \rightarrow (y, PersonBelongsToOrganization, x)$
	PersonBelongsToOrganization	$(x, PersonBelongsToOrganization, y) \rightarrow (y, OrganizationHiredPerson, x)$
Location	CapitalCityOfCountry CityLocatedInCountry CityLocatedInState StateHasCapital StateLocatedInCountry	$(x, CapitalCityOfCountry, y) \rightarrow (x, CityLocatedInCountry, y)$ $(x, StateHasCapital, y) \rightarrow (y, CityLocatedInState, x)$

n that represent the ground truth (assumption of induction). We must prove for any ground truth where $|\mathcal{W}| = n$, there exist an assignment of values to embedding vectors of size $n+1$ that represent this ground truth.

Let (h, r, t) be one of the n true facts. Consider a modified ground truth which is identical to the ground truth with n true facts, except that (h, r, t) is assigned false. The modified ground truth has $n-1$ true facts and based on the assumption of the induction, we can represent it using some embedding vectors of size n. Let $q = \langle \mathbf{h}^+, \mathbf{r}^+, \mathbf{t}^+ \rangle + \langle \mathbf{t}^-, \mathbf{r}^-, \mathbf{h}^- \rangle$. We add an element to the end of all embedding vectors and set it to 0. This increases the vector size to $n+1$ but does not change any scores. Then we set $\mathbf{h}$ to 1, $\mathbf{r}$ to 1 and $\mathbf{t}$ to $q+1$. This ensure this triple is true for the new vectors, and no other probability of triple is affected. ■

Theorem 5 (Subsumption) *SimplE$^+$ guarantees subsumption using an inequality constraints.*

Proof *Assume $\forall h, t \in \mathcal{E}^* : (h, r, t) \rightarrow (h, s, t)$ as a non-trivial subsumption. As legal entity embeddings in SimplE$^+$ have non-negative elements, by adding the element-wise inequality constraint $\mathbf{s} \geq \mathbf{r}$, we force $\mu(\mathbf{h}, \mathbf{s}, \mathbf{t}) \geq \mu(\mathbf{h}, \mathbf{r}, \mathbf{t})$ for all $h, t \in \mathcal{E}^*$ which is forcing the subsumption.* ■

Objective Function and Training

Given the function μ, that maps embeddings to the probability of a triple, ideally we would like to minimize the following regularized negative log-likelihood function:

$$\mathcal{L}(\{\mathbf{e}\}, \{\mathbf{r}\}) = -\sum_{(h,r,t) \in \mathcal{W}} \log(\mu(\mathbf{h}, \mathbf{r}, \mathbf{t})) - \sum_{(h,r,t) \in \mathcal{W}^c} \log(1 - \mu(\mathbf{h}, \mathbf{r}, \mathbf{t})) + \Omega(\{\mathbf{e}\}, \{\mathbf{r}\})$$

where $\{\mathbf{e}\}$ represents entity embeddings, $\{\mathbf{r}\}$ represents relation embeddings and $\Omega(\{\mathbf{e}\}, \{\mathbf{r}\})$ is a regularization term. We use L2-regularization in our experiments. Optimizing $\mathcal{L}$ poses two challenges: **I**) we do not know the sets $\mathcal{W}$ and $\mathcal{W}^c$, as the purpose of KG completion is to produce these sets in the first place; **II**) the number of triples (specially in $\mathcal{W}^c$) is often too large, and for larger KGs exact calculation of these terms is often computationally unfeasible.

To address **I**, we use $\mathcal{K}$ as a surrogate for $\mathcal{W}$ and use its complement $\mathcal{K}^c = \mathcal{E} \times \mathcal{R} \times \mathcal{E} - \mathcal{K}$ instead of $\mathcal{W}^c$. To address the computational problem in **II**, we use stochastic optimization and follow the contrastive approach of Bordes et al. (2013): for each mini-batch of positive samples from KG, we produce a mini-batch of negative samples of the same size, by randomly "corrupting" the head or tail of the triple – *i.e.*, replacing it with a random entity.

Enforcing the subsumptions In order to enforce $\forall h, t \in \mathcal{E}^* : (h, r, t) \rightarrow (h, s, t)$, we add an equality constraint as $\mathbf{r} = \mathbf{s} - \delta_r$, where δ_r is a non-negative vector that specifies how r differs from s. We learn δ_r for all relations r that are in such a subsumption. This equality constraint guarantees the inequality constraint of Theorem 5.

Experimental Results

The objective of our empirical evaluations is two-fold: First, we want to see the practical implication of non-negativity constraints in terms of effectiveness of training and the quality of final results. Second, and more importantly, we would like to evaluate the practical benefit of incorporating prior knowledge in the form of subsumptions in sparse data regimes.

Datasets: We conducted experiments on four standard benchmarks: WN18, FB15K, Sport and Location. WN18 is a subset of WORDNET (Miller, 1995) and FB15K is a subset of FREEBASE (Bollacker et al., 2008). Sport and Location datasets are introduced by Wang et al. (2015), who created them using NELL (Mitchell et al., 2015). The relations in Sport and Location, along with the subsumptions, are listed in Table 3. Table 2 gives a summary of these datasets. For evaluation on WN18, FB15K, we split the existing triples in KG into the same train, validation, and test sets using the same split as (Bordes et al., 2013).

Evaluation Metrics: To evaluate different KG completion methods we need to use a train $\mathcal{N}$ and test $\mathcal{T}$ split, where $\mathcal{N} \cup \mathcal{T} = \mathcal{K}$. We use two evaluation metrics: HIT@T and Mean Reciprocal Rank (MRR). Both these measures rely on the *ranking* of a triple in the test set $(h, r, t) \in \mathcal{T}$, obtained by corrupting the head (or the tail) of the relation with $h' \neq h$ and estimating $\mu(h', r, t)$. An indicator for a good KG completion method is that (h, r, t) ranks high in the sorted list among corrupted triples.

Let $\text{rank}_h(h, r, t)$ be the ranking of $\mu(h, r, t)$ among all head-corrupted relations, and let $\text{rank}_t(h, r, t)$ denote a similar ranking with tail corruptions. MRR is the mean of the

Table 4: Results on WN18 and FB15K for SimplE and SimplE$^+$ without incorporating subsumptions.

	WN18					FB15K				
	MRR		Hit@			MRR		Hit@		
Model	Filter	Raw	1	3	10	Filter	Raw	1	3	10
ComplEx	0.941	0.587	0.936	0.945	0.947	0.692	0.242	0.599	0.759	0.840
SimplE	0.942	0.588	0.939	0.944	0.947	0.727	0.239	0.660	0.773	0.838
SimplE$^+$	0.937	0.575	0.936	0.938	0.939	0.725	0.240	0.658	0.770	0.841

reciprocal rank:

$$\text{MRR} \doteq \frac{1}{2 * |\mathcal{T}|} \sum_{(h,r,t)\epsilon\mathcal{T}} \frac{1}{\text{rank}_h(h,r,t)} + \frac{1}{\text{rank}_t(h,r,t)}$$

To provide a better metric, Bordes et al. (2013) suggest removing any corrupted relation that is in KG. We refer to the original definition of MRR as raw MRR and to Bordes et al. (2013)'s modified version as filtered MRR.

HIT@T measures the proportion of triples in $\mathcal{T}$ that rank among top t after corrupting both heads and tails.

Effect of Non-Negativity Constraints

Non-negativity has been a subject studied in various research fields. In many NLP-related tasks, non-negativity constraints are studies to learn more interpretable representations for words (Murphy, Talukdar, and Mitchell, 2012). In matrix factorization, non-negativity constraints are used to produce more coherent and independent factors (Lee and Seung, 1999). Ding et al. (2018) also proposed using non-negativity constraint to incorporate subsumption into ComplEx. We use the non-negativity constraint in SimplE$^+$ to enforce monotonousity of probabilities as dictated by subsumption. In order to get non-negativity constraint on the embedding of entities, we simply apply an element-wise non-linearity $\phi : \mathbb{R} \rightarrow \mathbb{R}^{\geq 0}$ before evaluation – that is we replace $\mu(\mathbf{h}, \mathbf{r}, \mathbf{t})$ with $\mu(\phi(\mathbf{h}), \mathbf{r}, \phi(\mathbf{t}))$.

Table 1 shows the result of SimplE$^+$ with for different choices of ϕ: **I**) exponential $\phi(x) = e^x$; **II**) logistic $\phi(x) = (1 + e^{-x})^{-1}$; and **III**) rectified linear unit (ReLU) $\phi(x) = \max(x, 0)$. ReLU outperforms other choices, and therefore moving forward we use ReLU for non-negativity constraints.

Next, we evaluate the effect of non-negativity constraint on the performance of the algorithm. Table 4 shows our result on WN18 and FB15K datasets. Note that this is effectively comparing SimplE$^+$with SimplE and ComplEx, without accommodating any subsumptions. As the results indicate, this constraint does not deteriorate the model's performance.

Sparse Relations

In this section, we study the scenario of learning relations that appear in few triples in the KG. In particular, we observe the behaviour of various methods as the amount of training triples varies. We train SimplE, SimplE$^+$, and logical inference on fractions of the Sport training set and test

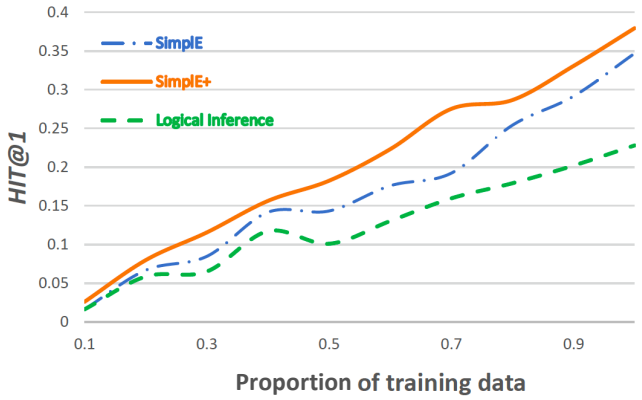

Figure 1: *hit@1 of SimplE, SimplE$^+$, and logical inference for different proportions of training data on Sport dataset*

them on the full test set. Logical inference refers to inferring new triples based only on the subsumptions.

Figure 1 shows the HIT@1 of the three methods when they are trained on different fractions (percentages) of the training data. According to Figure 1, when training data is scarce, logical inference performs better than (or on-par with) SimplE, as SimplE does not see enough triples to be able to learn meaningful embeddings. As the amount of training data increases, SimplE starts to outperform logical inference as it can better generalize to unseen cases than pure logical inference. The gap between these two methods becomes larger as the amount of training data increases. For all tested fractions, SimplE$^+$outperforms both SimplE and logical inference as it uses both the generalization power of SimplE and the inference power of logical rules.

In order to test the effect of incorporating taxonomical information on the number of epochs required for training to converge, we tested SimplE and SimplE$^+$on the Sport dataset with the same set of parameters and the same initialization and plotted the loss function for each epoch. The plot in Figure 2 shows that SimplE$^+$requires fewer epochs than SimplE to converge.

KGs with no Redundant Triples

Tensor factorization techniques rely on large amounts of annotated data. When background knowledge is available, we might expect a KG to not include redundant information. For instance if we have $(Paris, CapitalCityOfCountry, France)$ in a KG and

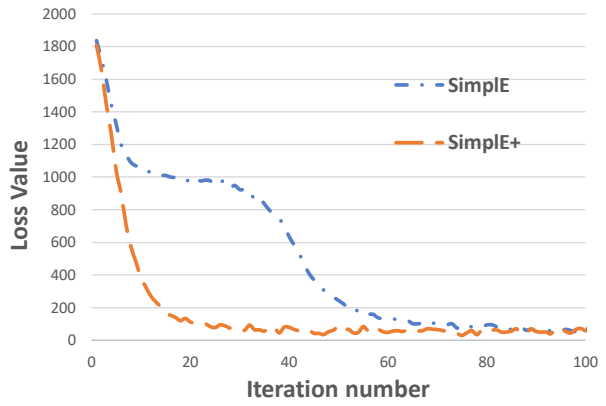

Figure 2: *Loss value at each epoch for SimplE and SimplE$^+$ on Sport dataset.*

we know $\forall h, t \in \mathcal{E}^* : (h, CapitalCityOfCountry, t) \rightarrow (h, CityLocatedInCountry, t)$, then the triple $(Paris, CityLocatedInCountry, France)$ is redundant. Similar to the experiment for incorporating background knowledge in Kazemi and Poole (2018c), we remove all redundant triples from the training set and compare SimplE with SimplE$^+$and logical inference. The obtained results in Table 5 demonstrate that SimplE$^+$outperforms SimplE and logical inference on both Sport and Location datasets with a large margin. As an example, SimplE$^+$gains almost 90 percent and 230 percent improvement over SimplE in terms of HIT@1 for Sport and Location datasets respectively. These results represent the clear advantage of SimplE$^+$over SimplE when background taxonomic information is available.

Related Work

Incorporating background knowledge in link prediction methods has been the focus of several studies. Here, we categorize these approaches emphasizing the shortcomings that are addressed in our work; see (Nickel et al., 2016) for a review of KG embedding methods.

Soft rules There is a large family of link prediction models based on soft first-order logic rules Richardson and Domingos (2006); De Raedt, Kimmig, and Toivonen (2007); Kazemi et al. (2014). While these models can be easily integrated with background taxonomic information, they typically cannot generalize to unseen cases beyond their rules. Exceptions include Fatemi, Kazemi, and Poole (2016); Kazemi and Poole (2018b) which combine (stacked layers of) soft rules with entity embeddings, but these models have only applied to property prediction. Approaches based on path-constrained random walks (e.g., Lao and Cohen (2010)) suffer from similar limitations as they have been shown to be a subset of probabilistic logic-based models Kazemi and Poole (2018a).

Augmentation by grounding of the rules The simplest way to incorporate a set of rules in the KG is to augment the KG with their groundings (Sedghi and Sabharwal, 2018) *before* learning the embedding. Demeester, Rocktäschel, and Riedel (2016) address the computational inefficiency of this approach through lifted rule injection. However, in addition to being inefficient, the the resulting model does not guarantee the subsumption in the completed KG.

Augmentation through post-processing A simple approach is to augment the KG *after* learning the embedding using an existing method (Wang et al., 2015; Wei et al., 2015). That is, as a post processing step we can modify the output of KG completion so as to satisfy the ontological constraints. The drawback of this approach is that the background knowledge does not help learn a better representation.

Regularized embeddings Rocktäschel, Singh, and Riedel (2015) regularize the learned embeddings using first-order logic rules. In this work, every logic rule is grounded based on observations and a differentiable term is added to the loss function for every grounding. For example, grounding the rule $\forall x : human(x) \rightarrow animal(x)$ would result in a very large number of loss terms to be added to the loss function in a large KG. This method as well as other approaches in this category (*e.g.*, Rocktäschel et al., 2014; Wang et al., 2015; Wang and Cohen, 2016) do not scale beyond a few entities and rules, because of the very large number of regularization terms added to the loss function (Demeester, Rocktäschel, and Riedel, 2016). Guo et al. (2018) proposed a methods for incorporating entailment into ComplEx called RUGE which models rules based on t-norm fuzzy logic, which imposes an independence assumption over the atoms. Such an independence assumption is not necessarily true, especially in the case of subsumption, e.g. in $human(x) \rightarrow animal(x)$ for which the left and the right part of the subsumption are strongly dependent. In addition to being inefficient, the resulting model of the regularized embedding approaches does not guarantee the subsumption

Table 5: Results on Sport and Location. Best results are in bold. MRR and Hit@n for n > 1 does not make sense for logical inference.

	Sport					Location				
	MRR		Hit@			MRR		Hit@		
Model	Filter	Raw	1	3	10	Filter	Raw	1	3	10
Logical inference	-	-	0.288	-	-	-	-	0.270	-	-
SimplE	0.230	0.174	0.184	0.234	0.324	0.190	0.189	0.130	0.210	0.315
SimplE$^+$	**0.404**	**0.337**	**0.349**	**0.440**	**0.508**	**0.440**	**0.434**	**0.430**	**0.440**	**0.450**

in the completed KG.

Constrained matrix factorization Several recent works incorporate background ontologies into the embeddings learned by matrix factorization (*e.g.*, Rocktäschel, Singh, and Riedel, 2015; Demeester, Rocktäschel, and Riedel, 2016). While these methods address the problems of the two categories above, they are inadequate due to the use of matrix factorization. Application of matrix factorization for KG completion (Riedel et al., 2013) learns a distinct embedding for each head-tail combination. In addition to its prohibitive memory requirement, since entities do not have their own embeddings, some regularities in the KG are ignored; for example this representation is oblivious to the fact that (h_i, r_k, t_j) and (h_l, r_m, t_j) share the same tail.

Constrained translation-based methods In translation-based methods, the relation between two entities is represented using an affine transformation, often in the form of translation. Most relevant to our work is KALE (Guo et al., 2016) that constrains the representation to accommodate logical rules, albeit after costly propositionalization. Several recent works show that a variety of existing translation-based methods are not fully expressive (Wang et al., 2017; Kazemi and Poole, 2018c), putting a severe limitation on the kinds of KGs that can be modeled using translation-based approaches.

Region based representation Gutiérrez-Basulto and Schockaert (2018) propose representing relations as convex regions in a $2k$-dimensional space, where k is the length of the entity embeddings. A relation between two embeddings is deemed true if the corresponding point is in the convex region of the relation. Although this framework allows Gutiérrez-Basulto and Schockaert (2018) to incorporate a subset of existential rules by restricting the convex regions of relations, they did not propose a practical method for learning and their method is restricted to a subset of existential rules.

Conclusion and Future Work

In this paper, we proposed SimplE$^+$, a fully expressive tensor factorization model for knowledge graph completion when background taxonomic information (in terms of subclasses and subproperties) is available. We showed that existing fully expressive models cannot provably respect subclass and subproperty information. Then we proved that by adding non-negativity constraints to entity embeddings of SimplE, a state-of-the-art tensor factorization approach, we can build a model that is not only fully expressive but also able to enforce subsumptions. Experimental results on benchmark KGs demonstrate that SimplE$^+$is simple yet effective. On our benchmarks, SimplE$^+$outperforms SimplE and offers a faster convergence rate when background taxonomic information is available. In future, we plan to extend SimplE$^+$to further incorporate ontological background information, and rules such as $\forall h, t \in \mathcal{E}^* : (h, r, t) \wedge (h, s, t) \rightarrow (h, p, t)$.

References

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 1247–1250. AcM.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, 2787–2795.

Carlson, A.; Betteridge, J.; Kisiel, B.; Settles, B.; Hruschka Jr, E. R.; and Mitchell, T. M. 2010. Toward an architecture for never-ending language learning. In *AAAI*, volume 5, 3. Atlanta.

De Raedt, L.; Kimmig, A.; and Toivonen, H. 2007. Problog: A probabilistic prolog and its application in link discovery.

Demeester, T.; Rocktäschel, T.; and Riedel, S. 2016. Lifted rule injection for relation embeddings. *arXiv preprint arXiv:1606.08359*.

Ding, B.; Wang, Q.; Wang, B.; and Guo, L. 2018. Improving knowledge graph embedding using simple constraints. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*.

Dong, X.; Gabrilovich, E.; Heitz, G.; Horn, W.; Lao, N.; Murphy, K.; Strohmann, T.; Sun, S.; and Zhang, W. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 601–610. ACM.

Fatemi, B.; Kazemi, S. M.; and Poole, D. 2016. A learning algorithm for relational logistic regression: Preliminary results. *arXiv preprint arXiv:1606.08531*.

Getoor, L., and Taskar, B. 2007. *Introduction to statistical relational learning*, volume 1. MIT press Cambridge.

Grover, A., and Leskovec, J. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 855–864. ACM.

Guo, S.; Wang, Q.; Wang, L.; Wang, B.; and Guo, L. 2016. Jointly embedding knowledge graphs and logical rules. In *EMNLP*, 192–202.

Guo, S.; Wang, Q.; Wang, L.; Wang, B.; and Guo, L. 2018. Knowledge graph embedding with iterative guidance from soft rules. In *AAAI*.

Gutiérrez-Basulto, V., and Schockaert, S. 2018. From knowledge graph embedding to ontology embedding: Region based representations of relational structures. *arXiv preprint arXiv:1805.10461*.

Hitzler, P.; Krötzsch, M.; Parsia, B.; Patel-Schneider, P. F.; and Rudolph, S., eds. 2012. *OWL 2 Web Ontology Language Primer (Second Edition)*. W3C Recommendation 11 December 2012.

Hoff, P. D.; Raftery, A. E.; and Handcock, M. S. 2002. Latent space approaches to social network analysis. *J. of the American Statistical association* 97(460):1090–1098.

Kazemi, S. M., and Poole, D. 2018a. Bridging weighted rules and graph random walks for statistical relational models. *Frontiers in Robotics and AI* 5:8.

Kazemi, S. M., and Poole, D. 2018b. Relnn: a deep neural model for relational learning.

Kazemi, S. M., and Poole, D. 2018c. Simple embedding for link prediction in knowledge graphs. In *NIPS*.

Kazemi, S. M.; Buchman, D.; Kersting, K.; Natarajan, S.; and Poole, D. 2014. Relational logistic regression. In *KR*. Vienna.

Lao, N., and Cohen, W. W. 2010. Relational retrieval using a combination of path-constrained random walks. *Machine learning* 81(1):53–67.

Lee, D. D., and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*, 3111–3119.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Mitchell, T.; Cohen, W.; Hruschka, E.; Talukdar, P.; Betteridge, J.; Carlson, A.; Dalvi, B.; Gardner, M.; Kisiel, B.; Krishnamurthy, J.; Lao, N.; Mazaitis, K.; Mohamed, T.; Nakashole, N.; Platanios, E.; Ritter, A.; Samadi, M.; Settles, B.; Wang, R.; Wijaya, D.; Gupta, A.; Chen, X.; Saparov, A.; Greaves, M.; and Welling, J. 2015. Never-ending learning. In *AAAI*.

Murphy, B.; Talukdar, P.; and Mitchell, T. 2012. Learning effective and interpretable semantic models using non-negative sparse embedding. *Proceedings of COLING 2012* 1933–1950.

Nguyen, D. Q. 2017. An overview of embedding models of entities and relationships for knowledge base completion. *arXiv preprint arXiv:1703.08098*.

Nickel, M.; Murphy, K.; Tresp, V.; and Gabrilovich, E. 2016. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE* 104(1):11–33.

Nickel, M.; Tresp, V.; and Kriegel, H.-P. 2012. Factorizing yago: scalable machine learning for linked data. In *Proceedings of the 21st international conference on World Wide Web*, 271–280. ACM.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*, 1532–1543.

Perozzi, B.; Al-Rfou, R.; and Skiena, S. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 701–710. ACM.

Raedt, L. D.; Kersting, K.; Natarajan, S.; and Poole, D. 2016. Statistical relational artificial intelligence: Logic, probability, and computation. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 10(2):1–189.

Richardson, M., and Domingos, P. 2006. Markov logic networks. *Machine learning* 62(1-2):107–136.

Riedel, S.; Yao, L.; McCallum, A.; and Marlin, B. M. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of ACL: Human Language Technologies*, 74–84.

Rocktäschel, T.; Bošnjak, M.; Singh, S.; and Riedel, S. 2014. Low-dimensional embeddings of logic. In *Proceedings of the ACL 2014 Workshop on Semantic Parsing*, 45–49.

Rocktäschel, T.; Singh, S.; and Riedel, S. 2015. Injecting logical background knowledge into embeddings for relation extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the ACL: Human Language Technologies*, 1119–1129.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. In *NIPS*, 4967–4976.

Sedghi, H., and Sabharwal, A. 2018. Knowledge completion for generics using guided tensor factorization. *Transactions of the Association of Computational Linguistics* 6:197–210.

Shoham, Y. 2016. Why knowledge representation matters. *Communications of the ACM* 59(1):47–49.

Socher, R.; Chen, D.; Manning, C. D.; and Ng, A. 2013. Reasoning with neural tensor networks for knowledge base completion. In *NIPS*, 926–934.

Trouillon, T.; Welbl, J.; Riedel, S.; Gaussier, É.; and Bouchard, G. 2016. Complex embeddings for simple link prediction. In *ICML*, 2071–2080.

Trouillon, T.; Dance, C. R.; Gaussier, É.; Welbl, J.; Riedel, S.; and Bouchard, G. 2017. Knowledge graph completion via complex tensor factorization. *JML* 18(1):4735–4772.

Wang, W. Y., and Cohen, W. W. 2016. Learning first-order logic embeddings via matrix factorization. In *IJCAI*, 2132–2138.

Wang, Q.; Wang, B.; Guo, L.; et al. 2015. Knowledge base completion using embeddings and rules. In *IJCAI*, 1859–1866.

Wang, Q.; Mao, Z.; Wang, B.; and Guo, L. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering* 29(12):2724–2743.

Wei, Z.; Zhao, J.; Liu, K.; Qi, Z.; Sun, Z.; and Tian, G. 2015. Large-scale knowledge base completion: Inferring via grounding network sampling over selected instances. In *ICKM*, 1331–1340. ACM.

Yang, B.; Yih, W.-t.; He, X.; Gao, J.; and Deng, L. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.