Universal Representation Learning of Knowledge Bases by Jointly Embedding Instances and Ontological Concepts

Junheng Hao, Muhao Chen, Wenchao Yu, Yizhou Sun, Wei Wang
University of California, Los Angeles
{jhao,yzsun,weiwang}@cs.ucla.edu,{muhaochen,yuwenchao}@ucla.edu

ABSTRACT

Many large-scale knowledge bases simultaneously represent two views of knowledge graphs (KGs): an *ontology view* for abstract and commonsense concepts, and an *instance view* for specific entities that are instantiated from ontological concepts. Existing KG embedding models, however, merely focus on representing one of the two views alone. In this paper, we propose a novel two-view KG embedding model, JOIE, with the goal to produce better knowledge embedding and enable new applications that rely on multi-view knowledge. JOIE employs both cross-view and intra-view modeling that learn on multiple facets of the knowledge base. The cross-view association model is learned to bridge the embeddings of ontological concepts and their corresponding instance-view entities. The intra-view models are trained to capture the structured knowledge of instance and ontology views in separate embedding spaces, with a hierarchy-aware encoding technique enabled for ontologies with hierarchies. We explore multiple representation techniques for the two model components and investigate with nine variants of JOIE. Our model is trained on large-scale knowledge bases that consist of massive instances and their corresponding ontological concepts connected via a (small) set of cross-view links. Experimental results on public datasets show that the best variant of JOIE significantly outperforms previous models on instance-view triple prediction task as well as ontology population on ontology-view KG. In addition, our model successfully extends the use of KG embeddings to entity typing with promising performance.

CCS CONCEPTS

• **Computing methodologies** → **Knowledge representation and reasoning**; *Semantic networks*; *Ontology engineering*.

KEYWORDS

Knowledge Graph; Relational Embeddings; Ontology Learning

ACM Reference Format:
Junheng Hao, Muhao Chen, Wenchao Yu, Yizhou Sun, Wei Wang. 2019. Universal Representation Learning of Knowledge Bases by Jointly Embedding Instances and Ontological Concepts. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3292500.3330838

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330838

1 INTRODUCTION

Knowledge bases (KBs), such as DBpedia [18], YAGO [23] and ConceptNet [34], have incorporated large-scale multi-relational data and motivated many knowledge-driven applications. These KBs store knowledge graphs (KGs) that can be categorized as two views: (i) the **instance-view knowledge graphs** that contain **relations** between specific **entities** in triples (for example, "*Barack Obama*", "*isPoliticianOf*", "*United States*") and (ii) the **ontology-view knowledge graphs** that constitute semantic **meta-relations** of abstract **concepts** (such as "*polication*", "*is leader of*", "*city*"). In addition, KBs also provide **cross-view** links that connect ontological concepts and instances, denoting whether an instance is an instantiation from a specific concept. Figure 1 shows a snapshot of such a KB.

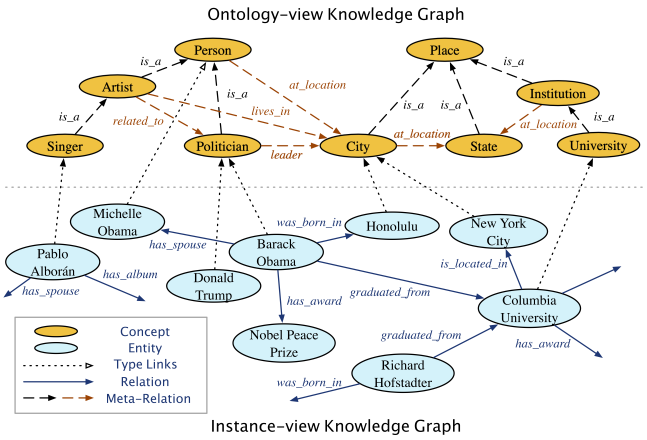


Figure 1: An example of two-view KB. Regular meta-relations and hierarchical meta-relations are denoted as orange and black dashed lines respectively in the ontology view.

In the past decade, KG embedding models have been widely investigated. These models, which encode KG structures into low-dimensional embedding spaces, are vital to capturing the latent semantic relations of entities and concepts, and support relational inferences in the form of vector algebra. They have provided an efficient and systematic solution to various knowledge-driven machine learning tasks, such as relation extraction [40], question answering [3], dialogues agents [14], knowledge alignment [6] and visual semantic labeling [10]. Existing embedding models, however, are limited to only one single view, either on the instance-view graph [2, 25, 42] or on the ontology-view graph [5, 13].

Learning to represent a KB from both views will no doubt provide more comprehensive insights. On one hand, instance embeddings provide detailed and rich information for their corresponding ontological concepts. For example, by observing many individual musicians, the embedding of its corresponding concept "*Musician*" can be largely determined. On the other hand, a concept embedding provides a high-level summary of its instances, which is extremely helpful when an instance is rarely observed. For example, for a musician who has few relational facts in the instance-view graph, we can still tell his or her rough position in instance embedding space because he or she should not be far away from other musicians.

In this paper, we propose to jointly embed the instance-view graph and the ontology-view graph, by leveraging (1) triples in both graphs and (2) cross-view links that connect the two graphs. It is a non-trivial task to effectively combine representation learning techniques on both views of a KB together, which faces the following challenges: (1) the vocabularies of entities and concepts, as well as relations and meta-relations, are disjoint but semantically related in these two views of the KB, and the semantic mappings from entities to concepts and from relations to meta-relations are complicated and difficult to be precisely captured by any current embedding models; (2) the known cross-view links often inadequately cover a vast number of entities, which leads to insufficient information to align both views of the KB, and curtails discovering new cross-view links; (3) the scales and topological structures are also largely different in the two views, where the ontological views are often sparser, provide fewer types of relations, and form hierarchical substructures, and the instance view is much larger and with much more relation types.

To address the above issues, we propose a novel KG embedding model named JOIE, which jointly encodes both the ontology and instance views of a KB. JOIE contains two components. First, a *cross-view association model* is designed to associate the instance embedding to its corresponding concept embedding. Second, the *intra-view embedding model* characterizes the relational facts of ontology and instance views in two separate embedding spaces. For the cross-view association model, we explore two techniques to capture the cross-view links. The *cross-view grouping* technique assumes that the two views can be forced into the same embedding space, while the *cross-view transformation* technique enables non-linear transformations from the instance embedding space to the ontology embedding space. As for the intra-view embedding model, in particular, we use three state-of-the-art translational or similarity-based relational embedding techniques to capture the multi-relational structures of each view. Additionally, for some KBs where ontologies contain hierarchical substructures, we employ a *hierarchy-aware* embedding technique based on intra-view non-linear transformations to preserve such substructures. Accordingly, we investigate nine variants of JOIE and evaluate these models on two tasks: the triple completion task and the entity typing task. Experimental results on the triple completion task confirm the effectiveness of JOIE for populating knowledge in both ontology and instance-view KGs, which has significantly outperformed various baseline models. The results on the entity typing task show that our model is competent in discovering cross-view links to align the ontology-view and the instance-view KGs.

2 RELATED WORK

To the best of our knowledge, there is no previous work on learning to embed two-view knowledge of a KB. We discuss the following three lines of research work that are closely relevant to this paper.

Knowledge Graph Embeddings. Recent work has put extensive efforts in learning instance-view KG embeddings. Given triples (h, r, t), where r represents the relation between the head entity h and the tail entity t, the key of KG embeddings is to design a plausibility scoring function $f_r(\mathbf{h}, \mathbf{t})$ as the optimization objective ($\mathbf{h}$ and $\mathbf{t}$ are embeddings of h and t). A recent survey [38] categorizes the majority of KG embedding models into translational models and similarity-based models. The representative translational model, TransE [2], adopts the score function $f_r(\mathbf{h}, \mathbf{t}) = -||\mathbf{h} + \mathbf{r} - \mathbf{t}||$ to capture the relation as a translation vector $\mathbf{r}$ between two entity vectors. Follow-ups of TransE typically vary the translation processes in different forms of relation-specific spaces, so as to improve the performance of triple completion. Examples include TransH [40], TransR [19], TransD [15] and TransA [16], etc. As for the similarity-based models, DistMult [42] associates related entities using Hadamard product of embeddings, and HolE [25] substitutes Hadamard product with circular correlation to improve the encoding of asymmetric relations, and achieves the state-of-the-art performance in KG completion. ComplEx [37] migrates DistMult in a complex space and offers comparable performance. Besides, there are other forms of models, including tensor-factorization-based RESCAL [26], and neural models NTN [33] and ConvE [8]. These approaches also achieve comparable performances on triple completion tasks at the cost of high model complexity.

It is noteworthy that a few approaches have been proposed to incorporate complex type information of entities into above KG embedding techniques [17, 21, 22, 41], from which our settings are substantially different in two perspectives: (i) These studies utilize the proximity of entity types to strengthen the learning of instance-level entity similarity, while do not capture the semantic relations between such types; (ii) They mostly focus on improving instance-view triple completion, but do not leverage instance-view knowledge to improve ontology population, nor support cross-view association to bridge instances and ontological concepts. Another related branch on leveraging logic rules [9, 12, 31] requires additional information that typically is not provided in two-view KBs.

Multi-graph Embeddings for KGs. More recent studies have extended embedding models to bridge multiple KG structures, typically for multilingual learning. MTransE [6] thereof, jointly learns a transformation across two separate translational embedding spaces, which can be adopted to our problem. However, since this multilingual learning approach partly relies on similar structures of KGs, it unsurprisingly falls short of capturing the associations between the two views of KB with disjoint vocabularies and different topologies, as we show in the experiments. Later extensions of this model family, such as KDCoE [4] and JAPE [35], require additional information of literal descriptions [4] and numerical attributes of entities [35] that are typically not available in the ontology views of the KB. Other models depend on the use of neural machine translation [27], causal reasoning [43] and bootstrapping of strictly 1-to-1 matching of inter-graph entities [36, 44] that do not apply to the nature of our corpora and tasks.

Ontology Population. Traditional ontology population is mostly based on extensive manual efforts, or requires large annotated text corpora for the mining of the meta-relation facts [7, 11, 24, 39]. These previous approaches rely on intractable parsing or human efforts, which generate massive relation facts that are subject to frequent conflicts [28]. A few studies extend embedding techniques to general cross-domain ontologies like ConceptNet. Examples of such include On2Vec [5] that extends translational embeddings to capture the relational properties and hierarchies of ontological relations, and Gutiérrez-Basulto and Schockaert [13] propose to learn second-order proximity of concepts by combining chained logic rules with ontology embeddings. This shows the benefits of KG embeddings on predicting relational facts for ontology population, while we argue that such a task can be simultaneously enhanced with the characterization of the instance knowledge.

3 MODELING

In this section, we introduce our proposed model JOIE, which jointly embed entities and concepts using two model components: *cross-view association model* and *intra-view model*. We start with the formalization of two-view knowledge bases.

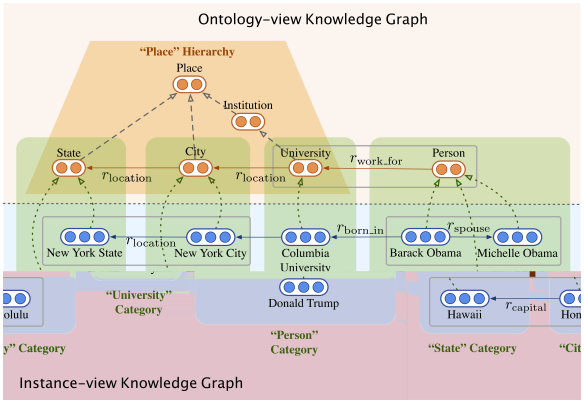


Figure 2: JOIE learns two aspects of a KB. The cross-view association model learns embeddings from cross-view links (dash arrows in green "category" box). The default intra-view model learns embeddings from triples (grey box) in each view; Besides, hierarchy-aware intra-view models the meta-relation facts that form hierarchies in the ontology (orange "Hierarchy" trapezoid).

3.1 Formalization of Knowledge Bases

In a KB, we use $\mathcal{G}_I$ and $\mathcal{G}_O$ to denote the instance-view KG and ontology-view KG respectively. The instance-view KG is denoted as $\mathcal{G}_I$, which is formed with $\mathcal{E}$, the set of entities, and $\mathcal{R}_I$, the set of relations. The set of concepts and meta-relations in the ontology-view graph $\mathcal{G}_O$ are similarly denoted as $\mathcal{C}$ and $\mathcal{R}_O$ respectively. Note that $\mathcal{E}$ and $\mathcal{C}$ (or $\mathcal{R}_I$ and $\mathcal{R}_O$) are disjoint sets. $(h^{(I)}, r^{(I)}, t^{(I)}) \in \mathcal{G}_I$ and $(h^{(O)}, r^{(O)}, t^{(O)}) \in \mathcal{G}_O$ denote triples in the instance-view KG and the ontology-view KG respectively, such that $h^{(I)}, t^{(I)} \in \mathcal{E}, h^{(O)}, t^{(O)} \in \mathcal{C}, r^{(I)} \in \mathcal{R}_I$, and $r^{(O)} \in \mathcal{R}_O$. Specifically, for each view in the KB, a dedicated low-dimensional space is assigned to embed nodes and edges. Boldfaced $\mathbf{h}^{(I)}, \mathbf{t}^{(I)}, \mathbf{r}^{(I)}$ represent the embedding vectors of head entity $h^{(I)}$, tail entity $t^{(I)}$ and relation $r^{(I)}$ in instance-view triples. Similarly, $\mathbf{h}^{(O)}, \mathbf{t}^{(O)}$, and $\mathbf{r}^{(O)}$ denote the embedding vectors for the corresponding concepts and their meta-relation in the ontology-view graph. Besides the notations for two views, $\mathcal{S}$ is used to denote the set of known cross-view links in the KB, which contains associations between instances and concepts such as *"type_of"*. We use $(e, c) \in \mathcal{S}$ to denote a link between $e \in \mathcal{E}$ and its corresponding concept $c \in \mathcal{C}$. For example, (e: Los Angeles International Airport, c: airport) denotes that "*Los Angeles International Airport*" is an instance of the concept "*airport*". Looking into the nature of the ontology view, we also have hierarchical substructures identified by *"subclass_of"* (or other similar meta-relations). That is, we can observe concept pairs $(c_l, c_h) \in \mathcal{T}$ that indicates a finer (more specific) concept belongs to a coarser (more general) concept. One aforementioned example is (c_l: singer, c_h: person).

Our model JOIE consists of two model components that learn embeddings from the two views: the cross-view association model enables the connection and information flow between the two views by capturing the instantiation of entities from corresponding concepts, and the intra-view model encodes the entities/concepts and relations/meta-relations on each view of the KB. The illustration of these model components for learning different aspects of the KB is shown in Figure 2. In the following subsections, we first discuss the cross-view association model and intra-view model for each view, then combine them into variants of proposed JOIE model.

3.2 Cross-view Association Model

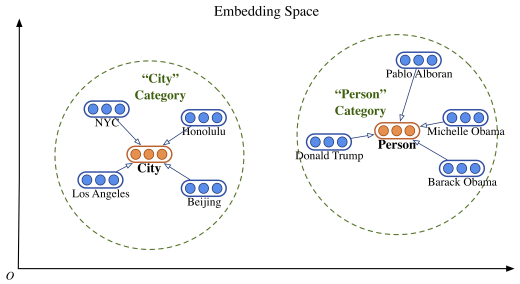
The goal of the cross-view association model is to capture the associations between the entity embedding space and the concept embedding space, based on the cross-view links in KBs, which will be our key contributions. We propose two techniques to model such associations: *Cross-view Grouping (CG)* and *Cross-view Transformation (CT)*. These two techniques are based on different assumptions and thus optimize different objective functions.

Cross-view Grouping (CG). The cross-view grouping method can be considered as grouping-based regularization, which assumes that the ontology-view KG and instance-view KG can be embedded into the same space, and forces any instance $e \in \mathcal{E}$ to be close to its corresponding concept $c \in \mathcal{C}$, as shown in Figure 3a. This requires the embedding dimensionalities for the instance-view and ontology-view graphs to be the same, i.e. $d = d_c = d_e$. Specifically, the categorical association loss for a given pair of cross-view link (e, c) is defined as the distance between the embeddings of e and c compared with margin γ^{CG}, and the loss is defined as,

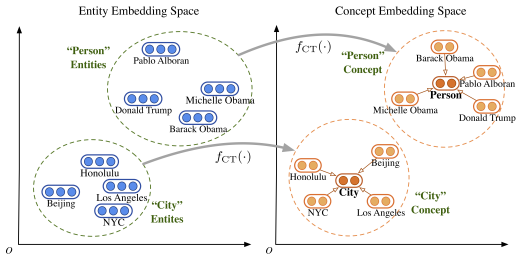
$$J^{\text{CG}}_{\text{Cross}} = \frac{1}{|\mathcal{S}|} \sum_{(e,c)\in\mathcal{S}} \left[||\mathbf{c} - \mathbf{e}||_2 - \gamma^{\text{CG}} \right]_+, \tag{1}$$

where $[x]_+$ is the positive part of the input x, i.e. $[x]_+ = \max\{x, 0\}$. This penalizes the case where the embedding of e falls out the γ^{CG}-radius[1] neighborhood centered at the embedding of c. CG has a strong clustering effect that makes entity embeddings close to their concept embeddings in the end.

[1]Typically, margin hyperparameter γ in the hinge loss can be chosen as 0.5 or 1 for different model settings. However, it is not a sensitive hyperparameter in our models.



(a) Cross-view Grouping (CG)



(b) Cross-view Transformation (CT)

Figure 3: Intuition of the cross-view association model: Cross-view Grouping (a); Cross-view Transformation (b).

Cross-view Transformation (CT). We also propose a cross-view transformation technique, which seeks to transform information between the entity embedding space and the concept space. Unlike CG that requires the two views to be embedded into the same space, the CT technique allows the two embedding spaces to be completely different from each other, which will be aligned together via a transformation, as shown in Figure 3b. In other words, after the transformation, an instance will be mapped to an embedding in the ontology-view space, which should be close to the embedding of its corresponding concept:

$$\mathbf{c} \leftarrow f_{\text{CT}}(\mathbf{e}), \forall (e, c) \in \mathcal{S}, \tag{2}$$

where $f_{\text{CT}}(\mathbf{e}) = \sigma(\mathbf{W}_{\text{ct}} \cdot \mathbf{e} + \mathbf{b}_{\text{ct}})$ is a non-linear affine transformation. $\mathbf{W}_{\text{ct}} \in \mathbb{R}^{d_2 \times d_1}$ thereof is a weight matrix and $\mathbf{b}_{\text{ct}}$ is a bias vector. $\sigma(\cdot)$ is a non-linear activation function, for which we adopt tanh.

Therefore, the total loss of the cross-view association model is formulated as Equation 3, which aggregates the CT objectives for all concepts involved in $\mathcal{S}$.

$$J_{\text{Cross}}^{\text{CT}} = \frac{1}{|\mathcal{S}|} \sum_{\substack{(e,c) \in \mathcal{S} \\ \wedge (e,c') \notin \mathcal{S}}} \left[\gamma^{\text{CT}} + ||\mathbf{c} - f_{\text{CT}}(\mathbf{e})||_2 - \left\| \mathbf{c}' - f_{\text{CT}}(\mathbf{e}) \right\|_2 \right]_+ \tag{3}$$

3.3 Intra-view Model

The aim of intra-view model is to preserve the original structural information in each view of the KB separately in two embedding spaces. Because of the different semantic meanings of relations in the instance view and meta-relations in the ontology view, it helps to give each view separate treatment rather than combining them into a single representation schema, improving the performance of downstream tasks, as shown in Section 4.2. In this section, we provide two intra-view model techniques for encoding heterogeneous and hierarchical graph structures.

Default Intra-view Model. To embed such a triple (h, r, t) in one KG, a score function $f(\mathbf{h}, \mathbf{r}, \mathbf{t})$ measures the plausibility of it. A higher score indicates a more plausible triple. Any triple embedding technique is applicable in our intra-view framework. In this paper, we adopt three representative techniques, i.e. translations [2], multiplications [42] and circular correlation [25]. The score functions of these techniques are given as follows.

$$\begin{aligned} f_{\text{TransE}}(\mathbf{h}, \mathbf{r}, \mathbf{t}) &= -||\mathbf{h} + \mathbf{r} - \mathbf{t}||_2 \\ f_{\text{Mult}}(\mathbf{h}, \mathbf{r}, \mathbf{t}) &= (\mathbf{h} \circ \mathbf{t}) \cdot \mathbf{r} \\ f_{\text{HolE}}(\mathbf{h}, \mathbf{r}, \mathbf{t}) &= (\mathbf{h} \star \mathbf{t}) \cdot \mathbf{r} \end{aligned} \tag{4}$$

where $\circ$ is the Hadamard product and $\cdot$ is the dot product. $\star : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ denotes circular correlation defined as $[\mathbf{a} \star \mathbf{b}]_k = \sum_{i=0}^{d} a_i b_{(k+i) \mod d}$.

To learn embeddings of all nodes in one graph $\mathcal{G}$, a hinge loss is minimized for all triples in the graph:

$$J_{\text{Intra}}^{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{\substack{(h,r,t) \in \mathcal{G} \\ \wedge (h',r,t') \notin \mathcal{G}}} \left[\gamma^{\mathcal{G}} + f(\mathbf{h}', \mathbf{r}, \mathbf{t}') - f(\mathbf{h}, \mathbf{r}, \mathbf{t}) \right]_+, \tag{5}$$

where $\gamma^{\mathcal{G}} > 0$ is a positive margin, and (h', r, t') is one sample from the set of corrupted triples which replace either head or tail entity and does not exist in $\mathcal{G}$.

The aforementioned techniques, losses and learning objectives for embedding graphs are naturally applicable for both instance-view graph and ontology-view graph. In the default intra-view model setting, for triples $(h^{(I)}, r^{(I)}, t^{(I)}) \in \mathcal{G}_I$ or $(h^{(O)}, r^{(O)}, t^{(O)}) \in \mathcal{G}_O$, we can compute $f_I(\mathbf{h}^{(I)}, \mathbf{r}^{(I)}, \mathbf{t}^{(I)})$ and $f_O(\mathbf{h}^{(O)}, \mathbf{r}^{(O)}, \mathbf{t}^{(O)})$ with the same techniques when optimizing $J_{\text{Intra}}^{\mathcal{G}_I}$ and $J_{\text{Intra}}^{\mathcal{G}_O}$. Combining the loss from instance-view and ontology-view graphs, the joint loss of the intra-view model is given as below,

$$J_{\text{Intra}} = J_{\text{Intra}}^{\mathcal{G}_I} + \alpha_1 \cdot J_{\text{Intra}}^{\mathcal{G}_O}, \tag{6}$$

where a positive hyperparameter α_1 weighs between the structural loss of the instance-view graph and ontology-view graph.

In JOIE deployed with the default Intra-view model, we employ the same triple encoding technique to represent both views of the KB. The purpose of doing so is to enforce the same paradigm of characterizing relational inferences in both views. It is noteworthy that there are other triple encoding techniques for KG embeddings, which can potentially be used in our intra-view model. Since exploring different triple encoding techniques is not the focus of our paper, we leave them as future work.

Hierarchy-Aware Intra-view Model for the Ontology. It is observed that the ontology view of some KBs form hierarchies, which is typically constituted by a meta-relation with the hierarchical property, such as "*subclass_of*" and "*is_a*" [18, 23]. We can define such meta-relation facts as $(c_l, r_{\text{meta}} = \text{"}subclass\_of\text{"}, c_h)$. For example, "*musician*" and "*singer*" belong to "*artist*" and "*artist*" is also subclass of "*person*". Such semantic ontological features requires additional modeling than other meta-relations. In other words, we further distinguish between meta-relations that form the ontology

hierarchy and those regular semantic relations (such as *"related_to"*) in our intra-view model.

To address this problem, we propose the hierarchy-aware (HA) intra-view model by extending a similar method to that of cross-view transformation as defined in Equation 2. Given concept pairs (c_l, c_h), we model such hierarchies into a non-linear transformation between coarser concepts and associated finer concepts by

$$g_{\text{HA}}(\mathbf{c}_h) = \sigma(\mathbf{W}_{\text{HA}} \cdot \mathbf{c}_l + \mathbf{b}_{\text{HA}}) \tag{7}$$

where $\mathbf{W}_{\text{HA}} \in \mathbb{R}^{d_2 \times d_2}$ and $\mathbf{b}_{\text{HA}} \in \mathbb{R}^{d_2}$ are defined similarly. Also, we use tanh function as $\sigma(\cdot)$ option. This will introduce a new loss term, ontology hierarchy loss inside the ontology view, which is similar to Equation 3,

$$J_{\text{Intra}}^{\text{HA}} = \frac{1}{|\mathcal{T}|} \sum_{\substack{(c_l, c_h) \in \mathcal{T} \\ \wedge (c_l, c_h') \notin \mathcal{T}}} \left[\gamma^{\text{HA}} + ||\mathbf{c}_h - g(\mathbf{c}_l)||_2 - \left\| \mathbf{c_h}' - g(\mathbf{c_l}) \right\|_2 \right]_+ \tag{8}$$

Therefore, the total training loss of the hierarchy-aware intra-view model for both views changes slightly to,

$$J_{\text{Intra}} = J_{\text{Intra}}^{\mathcal{G}_I} + \alpha_1 \cdot J_{\text{Intra}}^{\mathcal{G}_O \setminus \mathcal{T}} + \alpha_2 \cdot J_{\text{Intra}}^{\text{HA}} \tag{9}$$

where positive α_1 and α_2 are two weighing hyperparameters. In Equation 9, $J_{\text{Intra}}^{\mathcal{G}_O \setminus \mathcal{T}}$ refers to the loss of the default intra-view model that is only trained on triples with regular semantic relations. $J_{\text{Intra}}^{\text{HA}}$ is explicitly trained on the triples with meta-relations that form the ontology hierarchy, which is a major difference from Equation 6.

As the conclusion of this subsection, in JOIE, the basic assumption is that KGs have ontology hierarchy and rich semantic relational features compared to social or citation networks. JOIE is able to encode such KG properties in its model architecture. Note that we are also aware of the fact that there are more comprehensive properties of relations and meta-relations in the two views such as logical rules of relations and entity types. Incorporating such properties into the learning process is left as future work.

3.4 Joint Training on Two-View KBs

Combining the intra-view model and cross-view association model, JOIE minimizes the following joint loss function:

$$J = J_{\text{Intra}} + \omega \cdot J_{\text{Cross}}, \tag{10}$$

where $\omega > 0$ is positive hyperparameter that balances between J_{Intra} and J_{Cross}.

Instead of directly updating J, our implementation optimizes $J_{\text{Intra}}^{\mathcal{G}_I}$, $J_{\text{Intra}}^{\mathcal{G}_O}$ and J_{Cross} alternately. In detail, we optimize $\theta^{\text{new}} \leftarrow \theta^{\text{old}} - \eta \nabla J_{\text{Intra}}$ and $\theta^{\text{new}} \leftarrow \theta^{\text{old}} - (\omega\eta) \nabla J_{\text{Cross}}$ in successive steps within one epoch. η is the learning rate, and ω differentiates between the learning rates for intra-view and cross-view losses.

We use the AMSGrad optimizer [30] to optimize the joint loss function. We initialize vectors by drawing from a uniform distribution on the unit spherical surface, and initialize matrices using random orthogonal initialization [32]. During the training, we enforce the constraint that the L2 norm of all entity and concept vectors to be 1, in order to prevent them from shrinking to zero. This follows the setting by [2, 25, 40, 42]. Negative sampling is used on both intra-view model and cross-view association model with a ratio of 1 (number of negative samples per positive one). A hinge loss is applied for both models with all variants.

3.5 Variants of JOIE and Complexity

Without considering the HA technique, we have six variants of JOIE given two options of cross-view association models in Section 3.2 and three options of intra-view models in Section 3.3. For simplicity, we use the names of its components to denote specific variants of JOIE, such as "JOIE-TransE-CT" represents JOIE with the cross-view transformation and TransE-based default intra-view embeddings. In addition, we incorporate the hierarchy-aware intra-view model for the ontology view into cross-view transformation model[2], which produces three additional model variants denoted as JOIE-HATransE-CT, JOIE-HAMult-CT, and JOIE-HAHolE-CT.

The model complexity depends on the cross-view association model and intra-view model for learning two-view KBs. We denote n_e, n_c, n_r, n_m as the number of total entities, concepts, relations and meta-relations (typically $n_e \gg n_c$) and d_e, d_c as embedding dimensions ($d_e = d_c$ if CG is used). The model complexity of parameter sizes is $O(n_e d_e + n_c d_c)$ for all CG-based variants and $O(n_e d_e + n_c d_c + d_e d_c)$ for all CT-based variants. An additional parameter size of $O(d_c^2)$ is needed if the hierarchy-aware intra-view model applies. Because of $n \gg d_e$ (or d_c), the parameter complexity is approximately proportional to the number of entities and the model training runtime complexity is proportional to the number of triples in the KG. For the task of triple completion in the KG, the time complexity for all variants is $O(n_e d_e)$ for the instance-view graph or $O(n_c d_c)$ for the ontology-view graph. To process each prediction case in the entity typing task, the time complexity is $O(n_c d_e)$ for CG and $O(n_c d_c d_e)$ for CT. Details about each task are curated in Section 4.2 and 4.3.

4 EXPERIMENTS

In this section, we evaluate JOIE with two groups of tasks: the triple completion task (Section 4.2) on both instance-view and ontology-view KGs and the entity typing task (Section 4.3) to bridge two views of the KB. Besides, we provide a case study in Section 4.4 on ontology population and long-tail entity typing. We also present hyperparameter study, effects of cross-view sufficiency and negative samples in Appendix A.

4.1 Datasets

To the best of our knowledge, existing datasets for KG embeddings consider only an instance view (e.g. FB15k [2]) or an ontology view (e.g. WN18 [1]). Hence, we prepare two new datasets: *YAGO26K-906* and *DB111K-174*, which are extracted from YAGO [23] and DBpedia [18] respectively. The detailed dataset construction process is described in Appendix B.

Table 1 provides the statistics of both datasets. Normally, the instance-view KG is significantly larger than the ontology-view graph. Also, we notice that the two KBs are different in the density of type links, i.e., DB111K-174 has a much higher entity-to-concept

[2]We later show in the experiments that CT-based variants consistently outperform CG-based variants and thus we only apply HA intra-view model settings to CT-based model variants.

Table 1: Statistics of datasets.

Dataset	Instance Graph $\mathcal{G}_I$			Ontology Graph $\mathcal{G}_O$			Type Links $\mathcal{S}$
	#Entities	#Relations	#Triples	#Concepts	#Meta-relations	#Triples	
YAGO26K-906	26,078	34	390,738	906	30	8,962	9,962
DB111K-174	111,762	305	863,643	174	20	763	99,748

ratio (643.4) than YAGO26K-906 (28.7). Datasets are available at `https://github.com/JunhengH/joie-kdd19`.

4.2 KG Triple Completion

The objective of triple completion is to construct the missing relation facts in a KG structure, which directly tests the quality of learned embeddings. In our experiment, this task spans into two sub-tasks for instance-view KG completion and ontology population. We perform the sub-tasks on both datasets with all JOIE variants compared with baseline models.

Evaluation Protocol First, we separate the instance-view triples into training set $\mathcal{G}_I^{\text{train}}$, validation set $\mathcal{G}_I^{\text{valid}}$ and test set $\mathcal{G}_I^{\text{test}}$, as well as separate similarly the ontology-view triples to $\mathcal{G}_O^{\text{train}}$, $\mathcal{G}_O^{\text{valid}}$ and $\mathcal{G}_O^{\text{test}}$. The percentage of the training, validation and test cases is approximately 85%, 5% and 10%, which is consistent to that of the widely used benchmark dataset [2] for instance-only KG embeddings. Each JOIE variant is trained on $\mathcal{G}_I^{\text{train}}$ and $\mathcal{G}_O^{\text{train}}$ triples along with all cross-view links $\mathcal{S}$. In the testing phase, given each query $(h, r, ?t)$, the plausibility scores $f(\mathbf{h}, \mathbf{r}, \tilde{\mathbf{t}})$ for triples formed with every $\tilde{t}$ in the test candidate set are computed and ranked by the intra-view model. We report three metrics for testing: mean reciprocal ranks (*MRR*), accuracy (*Hits*@1) and the proportion of correct answers ranked within the top 10 (*Hits*@10). All three metrics are preferred to be higher, so as to indicate better triple completion performance. Also, we adopt the filtered metrics as suggested in previous work which are aggregated based on the premise that the candidate space has excluded the triples that have been seen in the training set [2, 42].

As for the hyperparameters in training, we select the dimensionality d among $\{50, 100, 200, 300\}$ for concepts and entities, learning rate among $\{0.0005, 0.001, 0.01\}$, margin γ among $\{0.5, 1\}$. We also use different batch sizes according to the sizes of graphs. We fix the best configuration $d_e = 300, d_c = 50$ for CT and $d_e = d_c = 200$ for CG with $\alpha_1 = 2.5, \alpha_2 = 1.0$. We set $\gamma^{\mathcal{G}_I} = \gamma^{\mathcal{G}_O} = 0.5$ as the default for all TransE variants and $\gamma^{\mathcal{G}_I} = \gamma^{\mathcal{G}_O} = 1$ for all Mult and HolE variants. The training processes on all datasets and models are limited to 120 epochs.

Baselines We compare our model with TransE, DistMult and HolE as well as TransC [20]. We deploy the following variants of baselines: (i) We train these mono-graph models (TransE, DistMult and HolE) either on instance-view triples or ontology-view triples separately, denoted as (*base*) in Table 2; (ii) We also train TransE, DistMult and HolE based on all triples in both $\mathcal{G}_I^{\text{train}}$ and $\mathcal{G}_O^{\text{train}}$. For the second setting thereof, we incorporate cross-view links by adding one additional relation "*type_of*" to them, denoted as (*all*) in Table 2. (iii) TransC is trained on both views of a KB. TransC is a recent work that differentiates between the encoding process of concepts from instances. Note that TransC is equivalent to a simplified case of our JOIE-TransE-CG where no semantic meta relations in the ontology view are included. For that reason, TransC does not apply to the completion of the ontology view.

Results As reported in Table 2, we categorize the results into three different groups based on the intra-view models. Though three intra-view models have different capabilities, among all the baselines in same group, JOIE notably outperforms others by 6.8% on *MRR*, and 14.8% on *Hit*@10 on average. A significant improvement is achieved on the ontology-view of DB111K-174 with JOIE compared to concept embeddings trained with only ontology-view triples and even 10.4% average increment compared to "all"-setting baselines and 34.97% compared to "base"-setting baselines. These results indicate that JOIE has better ability to utilize information from the instance view to promote the triple completion in ontology view. Comparing different intra-view models, translation based models performs better than similarity based models on ontology population and instance-view KG completion on the DB111K-174dataset. This is because these graphs are sparse, and TransE is less hampered by the sparsity in comparison to the similarity-based techniques [29]. By applying the HA technique in the intra-view models with CT, the performance on instance-view triple completion is noticeably improved in most cases in comparison to the default intra-view CT-based models, especially in variants with translation and circular correlation based intra-view models.

Generally, JOIE provides an effective method to train two-view KB separately and both $\mathcal{G}_I$ and $\mathcal{G}_O$ benefit each other in learning better embeddings, producing promising results in the triple completion task.

4.3 Entity Typing

The entity typing task seeks to predict the associating concepts of certain given entities. Similar to the triple completion task, we rank all candidates and report the top-ranked answers for evaluation.

Evaluation Protocol We separate the cross-view links of each dataset into training and test sets with the ratio of 60% to 40%, denoted as $\mathcal{S}^{\text{train}}$ and $\mathcal{S}^{\text{test}}$ respectively. Each model is trained on the entire instance-view and ontology-view graphs with cross-view links $\mathcal{S}^{\text{train}}$. Hyperparameters are carried forward from the triple completion task, in order to evaluate under controlled variables. In the test phase, given a specific entity e_q, we rank the concepts based on their embedding distances from the projection of $\mathbf{e}_q$ in the concept embedding space. and calculate *MRR*, *Hit*@1 (i.e. accuracy) and *Hit*@3 on the test queries. We perform the entity typing task on both datasets with all JOIE variants compared with these baselines.

Baselines We compare with TransE, DistMult, HolE and MTransE. For baselines other than MTransE, we convert the cross-view links (e, c) to triples $(e, r_T = \text{"type\_of"}, c)$. Therefore, entity typing is equivalent to the triple completion task for these baseline models. For MTransE, we treat concepts and entities as different views (originally input as knowledge bases of two languages in [6]) in their model and test with distance-based ranking.

Table 2: Results of KG triple completion. H@1 and H@10 denote $Hit@1$ and $Hit@10$ respectively. For each group of model variants with the same intra-view model, the best results are bold-faced. The overall best results on each dataset are underscored.

Datasets	YAGO26K-906						DB111K-174					
Graphs	$\mathcal{G}_I$ KG Completion			$\mathcal{G}_O$ KG Completion			$\mathcal{G}_I$ KG Completion			$\mathcal{G}_O$ KG Completion		
Metrics	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10
TransE (base)	0.195	14.09	34.51	0.145	12.29	20.59	0.327	22.26	49.01	0.313	23.22	46.91
TransE (all)	0.187	13.73	35.05	0.189	14.72	24.36	0.318	22.70	48.12	0.539	47.90	61.84
TransC	0.252	15.71	37.79	–	–	–	0.359	24.83	49.31	–	–	–
JOIE-TransE-CG	0.264	16.38	35.45	0.189	11.16	29.44	0.394	27.75	51.20	0.598	53.84	71.79
JOIE-TransE-CT	0.292	**18.72**	44.14	0.240	14.49	33.47	0.443	32.10	67.89	**0.622**	**58.10**	72.97
JOIE-HATransE-CT	**0.306**	18.62	**51.72**	**0.263**	**16.72**	**38.46**	**0.473**	**33.79**	**71.37**	0.591	52.07	**79.65**
DistMult (base)	0.253	22.91	28.76	0.197	**17.72**	25.08	0.265	25.95	27.63	0.235	15.18	29.11
DistMult (all)	0.288	**24.06**	31.24	0.156	14.32	16.54	0.280	27.24	29.70	0.501	45.52	64.73
JOIE-Mult-CG	0.274	18.80	37.45	0.198	11.16	27.91	0.320	23.44	49.49	0.532	46.15	68.91
JOIE-Mult-CT	**0.309**	20.40	**46.15**	**0.207**	14.71	30.43	**0.404**	**26.55**	**60.86**	**0.563**	**50.50**	71.62
JOIE-HAMult-CT	0.296	19.39	45.48	0.202	13.72	**31.10**	0.369	24.82	55.86	0.521	38.46	**77.25**
HolE (base)	0.265	**25.90**	28.31	0.192	18.70	20.29	0.301	29.24	31.51	0.227	18.91	32.83
HolE (all)	0.252	24.22	26.56	0.138	11.29	14.43	0.295	28.70	30.32	0.432	38.80	56.05
JOIE-HolE-CG	0.253	18.75	34.11	0.167	13.04	22.33	0.361	24.13	46.15	0.469	41.89	62.16
JOIE-HolE-CT	0.313	20.40	47.80	0.229	**20.85**	28.42	0.425	29.09	66.88	**0.514**	**43.24**	69.23
JOIE-HAHolE-CT	**0.327**	22.42	**52.41**	**0.236**	16.72	**30.96**	**0.464**	**33.11**	**69.56**	0.503	40.80	**71.03**

Table 3: Results of entity typing.

Datasets	YAGO26K-906			DB111K-174		
Metrics	MRR	Acc.	Hit@3	MRR	Acc.	Hit@3
TransE	0.144	7.32	35.26	0.503	43.67	60.78
MTransE	0.689	60.87	77.64	0.672	59.87	81.32
JOIE-TransE-CG	0.829	72.63	93.35	0.828	70.58	95.11
JOIE-TransE-CT	0.843	75.31	93.18	0.846	74.41	94.53
JOIE-HATransE-CT	**0.897**	**85.60**	**95.91**	**0.857**	**75.55**	**95.91**
DistMult	0.411	36.07	55.32	0.551	49.83	68.01
JOIE-Mult-CG	0.762	62.62	87.82	0.764	60.83	91.80
JOIE-Mult-CT	0.805	70.83	89.25	**0.791**	65.30	**93.47**
JOIE-HAMult-CT	**0.865**	**81.63**	**91.83**	0.778	**69.38**	85.71
HolE	0.395	34.83	54.79	0.504	44.75	65.38
JOIE-HolE-CG	0.777	65.30	87.89	0.784	66.75	89.37
JOIE-HolE-CT	0.813	72.27	88.71	0.805	68.84	**91.22**
JOIE-HAHolE-CT	**0.888**	**83.67**	**93.87**	**0.808**	**72.51**	89.79

Results Results are reported in Table 3. All JOIE variants perform significantly better than the baselines. The best JOIE model, i.e. JOIE-TransE-CT, outperforms the best baseline model MTransE by 15.4% in terms of accuracy and 14.4% in terms of *MRR* on YAGO26K-906. The improvement on accuracy and *MRR* are 14.3% and 14.5% on DB111K-174 compared to MTransE. The results by other baselines confirm that the cross-view links, which apply to all entities and concepts, cannot be properly captured as a regular relation and requires a dedicated representation technique.

Considering different JOIE variants, our observation is that using translation based intra-view model and CT as the cross-view association model (JOIE-TransE-CT) is consistently better than other settings on both datasets. It has an average of 4.1% performance gain in *MRR* over JOIE-HolE-CT and JOIE-DistMult-CT, and an average of 2.17% performance gain in accuracy over the best of the rest variants (JOIE-TransE-CG). We believe that, compared with similarity-based intra-view models, translation based intra-view model better differentiates between different entities and different concepts in KGs with directed relations and meta-relations in the KB [29]. The results by CT-based model variants are generally better than those by CG-based ones. We believe this is due to two reasons: (i) CT allows the two embedding spaces have different dimensionalties, and hence better characterizes the ontology-view that is smaller and sparser than the instance view; (ii) As the topological structures of the two views may exhibit some inconsistency, CT adapts well and is less sensitive to such inconsistency than CG.

In terms of different intra-view models, it is also observed that HA intra-view model with CT settings can drastically enhance entity typing task and achieve the best performance especially for YAGO26K-906 with relatively rich ontology, which improves an average of 6.0% on *MRR* and 10.5% in accuracy compared with the default intra-view settings. The reason that the HA technique does not have similar effects on DB111K-174 is because DB111K-174 contains a small ontology with much smaller hierarchical structures[3]. Comparing the two datasets, our experiments show that, JOIE generally achieves similar accuracies and *MRR* scores on YAGO26K-906 and DB111K-174, but slightly better $Hit@3$ on DB111K-174 due to its smaller candidate space.

Our method opens up a new direction that the learned embedding may help guide labeling entities with unknown types. In Section 4.4 and Appendix A, we provide more experiments and insights on the benefits of representation learning with JOIE.

4.4 Case Study

In this section, we provide two case studies for ontology population and entity typing for long-tail entities.

Ontology Population By embedding the meta-relations and concepts in the ontology view, the triple completion process can already populate the ontology view with seen meta-relations, by answering the query like ("*Concert*","*Related to*",?*t*) in the KG completion task. Given the top answers of the query, we can reconstruct

[3]DB111K-174 contains 164 ontology-view triples for meta-relations with the hierarchical property, while YAGO26K-906 contains 1,411.

triples like ("*Concert*","*Related to*","*Ballet*") and ("*Concert*","*Related to*","*Musical*") with high confidence. However, this process does not resolve the zero-shot cases where some concepts may satisfy some meta-relations that have not pre-existed in the vocabulary of meta-relations. We cannot predict the potentially new meta-relation "is Politician of" directly with triple completion by answering the following query: ("*Office Holder*", ?r, "*Country*").

Our proposed JOIE provides a feasible solution by leveraging the cross-view association model that bridges the two views of the KG, and migrate proper instance-view relations to ontology-view meta-relations. This is realized by transforming the concept embeddings in the query to the entity embedding space, and selecting candidate relations from the instance-view. Considering the previous query ("*Office Holder*", ?r, "*Country*"), we first find the concept embeddings of "*Office Holder*" and "*Country*" (denoted as $\mathbf{c}_{\text{office}}$ and $\mathbf{c}_{\text{country}}$ respectively), and then transform them to the entity space. Specifically, for JOIE variants with translational intra-view model, we find the instance-view relations that are closest to $f^{\text{inv}}_{\text{CT}}(\mathbf{c}_{\text{country}}) - f^{\text{inv}}_{\text{CT}}(\mathbf{c}_{\text{office}})$. Figure 4 shows the PCA projections of the top 10 relation prediction results for this query. The top 3 relations are "*is Politician of*", "*is Leader of*" and "*is Citizen of*", which are all reasonable answers.

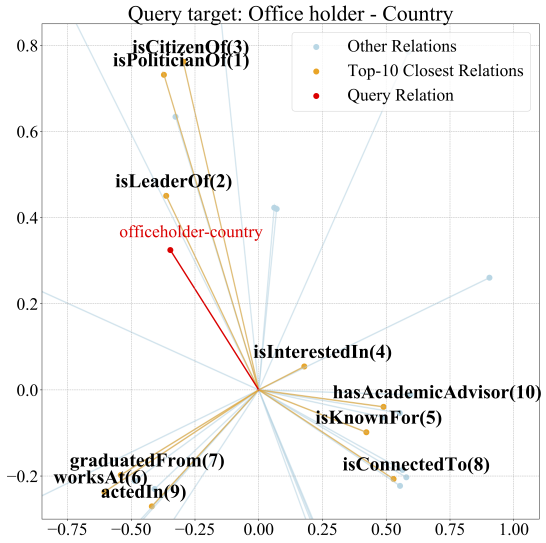


Figure 4: Examples of ontology population by finding the closest relations in the instance view for the query "Office Holder-Country". Top 10 predicted relations are plotted with their ranks.

Table 4 shows some examples of newly discovered meta-relation facts that have not pre-existed in the ontology views of the two datasets. Five predictions with the highest plausibility (smallest distance) are provided for each query from the ontology-view graph. From these top predictions, we observe that most populated ontology triples migrated from the instance view are meaningful.

Long-tail entity typing In KGs, the frequency of entities and relations often follow a long-tail distribution (Zipf's law). As shown in Figure 8a and Figure 8b (in Appendix B), both YAGO26K-906 and DB111K-174 discover such a property. Over 75% of total entities has less than 15 occurrences. Those long-tails entities, types and relations are difficult for representation learning algorithms to capture due to being few-shot in training cases.

Table 4: Examples of ontology population from JOIE-TransE-CT. Top 5 Populated Triples with smallest L2-norm distances are provided with reasonable answers bold-faced.

Query	Top 5 Populated Triples with distances
(scientist,?r, university)	scientist, ***graduated from***, university (0.499) scientist, ***isLeaderOf***, university (1.082) scientist, *isKnownFor*, university (1.098) scientist, *created*, university (1.119) scientist, ***livesIn***, university (1.141)
(boxer, ?r, club)	boxer, ***playsFor***, club (1.467) boxer, ***isAffiliatedTo***, club (1.474) boxer, ***worksAt***, club (1.479) boxer, *graduatedFrom*, club (1.497) boxer, *isConnectedTo*, club (1.552)
(TV station, ?r, country)	TV station, ***headquarter***, country (1.221) TV station, *parentOrganisation*, country (1.246) TV station, *appointer*, country (1.253) TV station, ***broadcastArea***, country (1.266) TV station, ***principalArea***, country (1.271)
(scientist, ?r, scientist)	scientist, *deputy*, scientist (0.204) scientist,***doctoralAdvisor***, scientist (0.218) scientist, ***doctoralStudent***, scientist (0.221) scientist, ***relative***, scientist (0.228) scientist, ***spouse***, scientist (0.230)

Table 5: Results of long-tail entities typing.

Datasets	YAGO26K-906			DB111K-174		
Metrics	MRR	Acc.	Hit@3	MRR	Acc.	Hit@3
DistMult	0.156	10.89	25.33	0.219	16.48	33.71
MTransE	0.526	46.45	67.25	0.505	46.67	64.36
JOIE-TransE-CG	0.708	59.97	79.80	0.741	64.45	83.05
JOIE-TransE-CT	0.737	62.05	82.60	0.758	66.35	83.80
JOIE-HATransE-CT	**0.802**	**69.66**	**87.75**	**0.760**	**67.34**	**89.79**

In this case study, we select the entities with considerably low frequency[4], which involve around 15%-30% of total entities in the instance view of the two KB datasets. Then, we evaluate the entity typing task for these long-tail entities. Table 5 shows the results by the best baselines (DistMult, MTransE) and a groups of our best JOIE variants. Similar to our previous observation, JOIE significantly outperforms other baselines. Compared with the results in Section 4.3, we observe the depletion of performance for all models, while JOIE variants only have an average of 12.5% decrease in *MRR* with CG models and 12.3% decrease in *MRR* with CT models while other baselines suffer over 20% on long-tail entity prediction. There is also an interesting observation that, for long-tails entities, smaller embeddings for both CG ($d_1 = d_2 = 100$) and CT ($d_1 = 100, d_2 = 50$) models are beneficial for associated concept prediction. We hypothesize that this is caused by overfitting on long-tail entities if high dimensionality is used for training without enough training data.

In Table 6, we include some examples of top 3 predicted categories of long-tail entities by DistMult, MTransE and JOIE (using JOIE-HATransE-CT variant) from DB111K-174, when the instance-view graph and ontology-view graph are relatively sparser. JOIE is

[4]In this experiment, we select entities in YAGO26K-906 which occurs less than 8 times and entities in DB111K-174 which occurs less than 3 times.

Table 6: Examples of long-tail entity typing. Top 3 predictions are provided with the correct type bold-faced.

Entity	Model	Top 3 Concept Prediction
Laurence Fishburne	DistMult	football team, club, team
	MTransE	writer, **person**, artist
	JOIE	**person**, artist, philosopher
Warangal City	DistMult	country, village,**city**
	MTransE	administrative region, **city**, settlement
	JOIE	**city**, town, country
Royal Victor -ian Order	DistMult	person, writer, administrative region
	MTransE	election, award, **order**
	JOIE	award, **order**, election

still able to make correct predictions of low-frequency entities while other baselines models can only output inaccurate predictions.

5 CONCLUSION AND FUTURE WORK

In this paper, we propose a novel model JOIE aiming to jointly embed real-world entities and ontological concepts. We characterize a two-view knowledge base. In the embedding space, our approach jointly captures both structured knowledge of each view, and cross-view links that bridges the two views. Extensive experiments on the tasks of KG completion and entity typing show that our model JOIE can successfully capture latent features from both views in KBs, and outperforms various state-of-the-art baselines.

We also point out future directions and improvements. Particularly, instead of optimizing structure loss with triples (first-order neighborhood) locally, we plan to adopt more complex embedding models which leverage information from higher order neighborhood, logic paths or even global knowledge graph structures. We also plan to explore the alignment on relations and meta-relations like entity-concept.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their insightful and constructive comments. This work was partially supported by NIH R01GM115833, U01HG008488, NSF DBI-1565137, DGE-1829071, NSF III-1705169, NSF Career Award 1741634 and Amazon Research Award.

REFERENCES

[1] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. 2014. A semantic matching energy function for learning with multi-relational data. *Machine Learning* 94, 2 (2014), 233–259.
[2] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*.
[3] Antoine Bordes, Jason Weston, and Nicolas Usunier. 2014. Open question answering with weakly supervised embedding models. In *ECML-PKDD*.
[4] Muhao Chen, Yingtao Tian, Kai-Wei Chang, Steven Skiena, and Carlo Zaniolo. 2018. Co-training Embeddings of Knowledge Graphs and Entity Descriptions for Cross-lingual Entity Alignment. *IJCAI* (2018).
[5] Muhao Chen, Yingtao Tian, Xuelu Chen, Zijun Xue, and Carlo Zaniolo. 2018. On2Vec: Embedding-based Relation Prediction for Ontology Population. In *SDM*.
[6] Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. (2017).
[7] Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *ACL*.
[8] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. *AAAI*.
[9] Jianfeng Du, Kunxun Qi, Hai Wan, Bo Peng, Shengbin Lu, and Yuming Shen. 2017. Enhancing Knowledge Graph Embedding from a Logical Perspective. In *JIST*. Springer, 232–247.
[10] Yuan Fang, Kingsley Kuan, Jie Lin, Cheston Tan, and Vijay Chandrasekhar. 2017. Object detection meets knowledge graphs. (2017).
[11] Claudio Giuliano and Alfio Gliozzo. 2008. Instance-based ontology population exploiting named-entity substitution. In *COLING*.
[12] Shu Guo, Quan Wang, Lihong Wang, Bin Wang, and Li Guo. 2016. Jointly embedding knowledge graphs and logical rules. In *EMNLP*.
[13] Vıctor Gutiérrez-Basulto and Steven Schockaert. 2018. From Knowledge Graph Embedding to Ontology Embedding? An Analysis of the Compatibility between Vector Space Representations and Rules. In *KR*.
[14] He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *ACL*.
[15] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *IJCNLP*.
[16] Yantao Jia, Yuanzhuo Wang, Hailun Lin, Xiaolong Jin, and Xueqi Cheng. 2016. Locally Adaptive Translation for Knowledge Graph Embedding.. In *AAAI*.
[17] Denis Krompaß, Stephan Baier, Volker Tresp, et al. 2015. Type-constrained representation learning in knowledge graphs. In *ISWC*.
[18] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* (2015).
[19] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion.. In *AAAI*.
[20] Xin Lv, Lei Hou, Juanzi Li, and Zhiyuan Liu. 2018. Differentiating Concepts and Instances for Knowledge Graph Embedding. In *EMNLP*.
[21] Jianxin Ma, Peng Cui, Xiao Wang, and Wenwu Zhu. 2018. Hierarchical taxonomy aware network embedding. In *KDD*. ACM.
[22] Shiheng Ma, Jianhui Ding, Weijia Jia, et al. 2017. Transt: Type-based multiple embedding representations for knowledge graph completion. In *ECML-PKDD*.
[23] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M Suchanek. 2015. Yago3: A knowledge base from multilingual Wikipedias. In *CIDR*.
[24] Hamid Mousavi, Maurizio Atzori, Shi Gao, and Carlo Zaniolo. 2014. Text-mining, structured queries, and knowledge management on web document corpora. *SIGMOD* (2014).
[25] Maximilian Nickel, Lorenzo Rosasco, Tomaso A Poggio, et al. 2016. Holographic Embeddings of Knowledge Graphs.. In *AAAI*.
[26] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A Three-Way Model for Collective Learning on Multi-Relational Data.. In *ICML*.
[27] Naoki Otani, Hirokazu Kiyomaru, Daisuke Kawahara, and Sadao Kurohashi. 2018. Cross-lingual Knowledge Projection Using Machine Translation and Target-side Knowledge Base Completion. In *COLING*.
[28] Jeff Pasternack and Dan Roth. 2013. Latent credibility analysis. In *WWW*. ACM.
[29] Jay Pujara, Eriq Augustine, and Lise Getoor. 2017. Sparsity and Noise: Where Knowledge Graph Embeddings Fall Short. In *EMNLP*.
[30] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2018. On the convergence of adam and beyond. In *ICLR*.
[31] Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting logical background knowledge into embeddings for relation extraction. In *NAACL-HLT*.
[32] Andrew M Saxe, James L McClelland, and Surya Ganguli. 2013. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *ICLR* (2013).
[33] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *NIPS*.
[34] Robert Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*.
[35] Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *ISWC*.
[36] Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping Entity Alignment with Knowledge Graph Embedding.. In *IJCAI*.
[37] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *ICML*.
[38] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE TKDE* (2017).
[39] Ting Wang, Yaoyong Li, Kalina Bontcheva, Hamish Cunningham, and Ji Wang. 2006. Automatic extraction of hierarchical relations from text. In *ESWC*.
[40] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes.. In *AAAI*.
[41] Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2016. Representation Learning of Knowledge Graphs with Hierarchical Types.. In *IJCAI*.
[42] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. *ICLR*.
[43] Jinyoung Yeo, Geungyu Wang, Hyunsouk Cho, Seungtaek Choi, and Seung-won Hwang. 2018. Machine-Translated Knowledge Transfer for Commonsense Causal Reasoning.. In *AAAI*.
[44] Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Iterative entity alignment via joint knowledge embeddings. In *AAAI*.

A ABLATION STUDY

In this section, we provide some insights on several critical factors that affect the performance of the model. These include the embedding dimensionality, sufficiency of cross-view links in training, and the effect of adopting negative sampling in cross-view association models.

A.1 Dimensionality

Dimensionality is a key hyperparameter that affects the quality of the obtained embeddings. Figure 5a shows the *MRR* of model variants with the CG-based cross-view association according to different embedding dimensions d. It is observed in Figure 5a that the performance of CG variants are generally improving from $d = 50$ to $d = 200$, however, after reaching the optimal $d_{\text{opt}} = 200$, *MRR* begins to drop at $d = 300$. Similarly we plot *MRR* scores for both dataset with CT model variants in Figure 5b.

We compare four different dimensionality settings of (d_1, d_2): $(100, 20)$,$(100, 50)$,$(300, 50)$ and $(300, 100)$[5]. Most of the JOIE variants achieve their best performance under the embedding setting $(d_1, d_2) = (300, 50)$ rather than $(d_1, d_2) = (300, 100)$ (except JOIE-Mult-CT on DB111K-174). The reason is that, JOIE set with low dimensionalities easily falls short of capturing latent features of entities and concepts, while too high dimensionalities lead to overfitting on the ontology view of KG, as well as inefficient training and prediction processes.

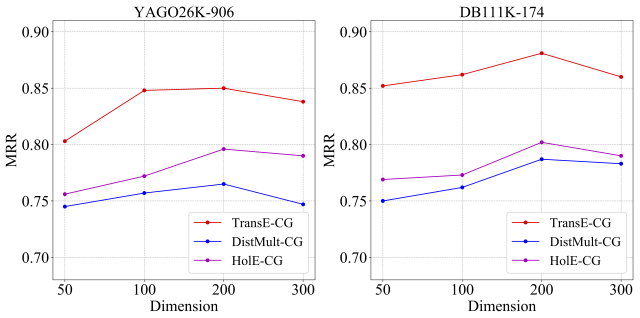
A.2 Sufficiency of Type Information

Cross-view links between the instance-view graph and the ontology-view graph are key components, which bridge and enable the information flow between two views to generate embeddings. We also investigate the influence of cross-view links and their sufficiency in training.

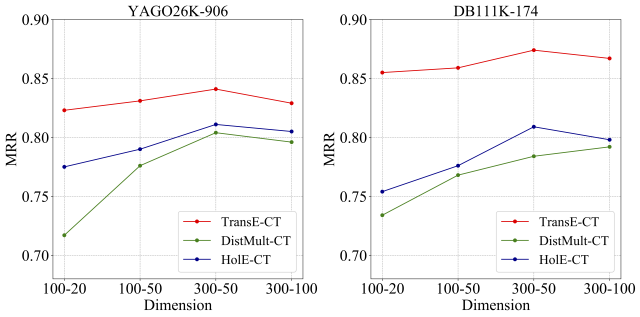
We define the train set ratio $\nu = \{0.2, 0.4, 0.6, 0.8\}$, which means the proportions of the cross-view links that are used for training JOIE. *MRR* score is reported in Figure 6a on YAGO26K-906 and Figure 6b on DB111K-174. As expected, when the proportion of cross-view links used for training increasing from 20% to 80%, the performance improves by 3.2% on YAGO26K-906 and by 2.9% on DB111K-174 in terms of *MRR*. It is noteworthy that JOIE trained with 20% cross-view links still outperforms MTransE trained with 60% cross-view links, which indicates that one advantage of JOIE is its outstanding generalization ability to other untyped entities, given limited knowledge on entity-concept pairs.

One interesting observation is that, when ν increases from 0.6 to 0.8, the performance of CG variants does not necessarily improve, while the performance of CT variants still has significant improvements. We hypothesize that this is because the strong clustering-based constraint in CG can be sensitive to even minor inconsistencies between the topological structures of the two KG views, giving too much supervision. CT, on the contrary, is more robust against the inconsistency between the two views. There is a trade-off between the robustness of CT and the efficiency of CG.

[5] $(d_1, d_2) = (100, 20)$ denotes that entities are embedded with $d_1 = 100$ dimensional vectors and concepts are embedded with $d_2 = 20$ dimensional vectors

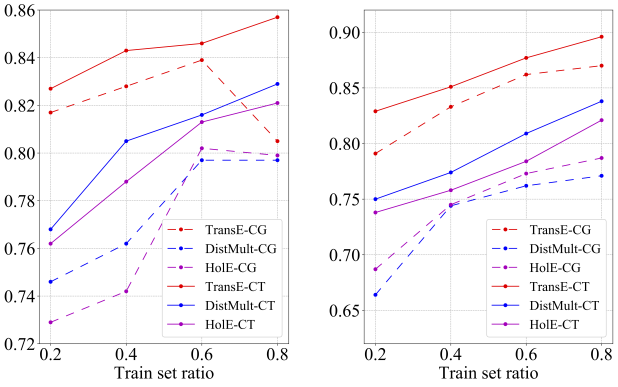


(a) Different dimensions with CG variants



(b) Different dimensions with CT variants

Figure 5: Performances of entity typing task on both datasets with different entity and concept embedding dimensionalities



(a) YAGO26K-906 (b) DB111K-174

Figure 6: The effect of training the model using different proportions of cross-view links on (a) YAGO26K-906 and (b) DB111K-174

A.3 Effects of Negative Sampling

Negative sampling is widely applied in the encoding process of a single KG structure [2, 42]. One interesting question is whether to use negative sampling for capturing the cross-view links between two structures, i.e. to provide corrupted entity-concept pairs such as ("Barack Obama","state"). We compare the results of entity typing task by JOIE variants with and without cross-view link negative

Table 7: Effects of negative sampling in type links

Datasets	YAGO26K-906		DB111K-174	
Setting	W/O NS	W/ NS	W/O NS	W/ NS
JOIE-TransE-CG	0.657	0.805	0.815	0.864
JOIE-Mult-CG	0.627	0.762	0.761	0.797
JOIE-HolE-CG	0.682	0.777	0.783	0.815
JOIE-TransE-CT	0.501	0.847	0.667	0.883
JOIE-Mult-CT	0.490	0.829	0.494	0.811
JOIE-HolE-CT	0.508	0.821	0.560	0.821

samples in Table 7. It is our finding that there is a significant performance drop if negative sampling is disabled in CT, while negative sampling has less effect on CG. We hypothesize that the difference is attributed to the fact that strong clustering-based constraint of CG is already effective in separating irrelevant concepts.

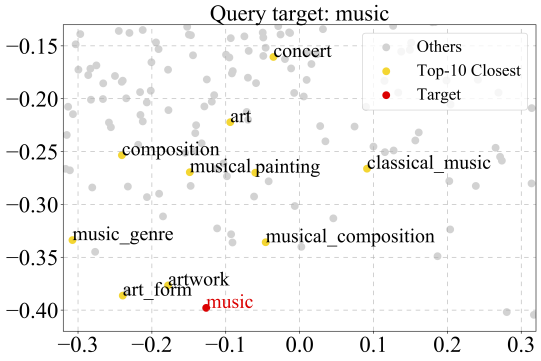
We show the effects of negative sampling by visualizing the results of one query, which are plotted as PCA projections in Figure 7. For the displayed query which targets at the concept "music", we plot the 10 nearest neighbors of concepts. Although related concepts such as "classic music", "concert" and "artist movement" still stay close by "music" in both settings, other irrelevant concepts including "decoration" and "architect" intercept in JOIE-TransE-CT without negative sampling. We find such phenomenon frequently exist in the JOIE embeddings trained without negative sampling, which no-doubt impairs the performance of the entity typing task.

B DATASETS

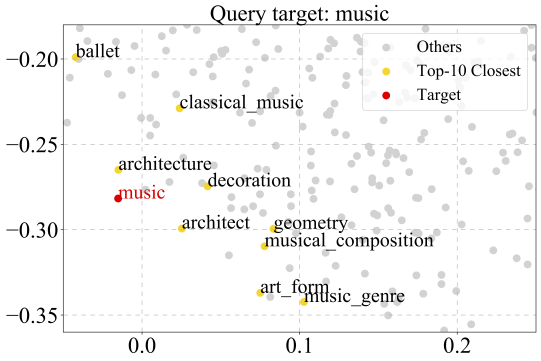
We use YAGO26K-906 and DB111K-174, which are extracted from the connected subsets of YAGO [23] and DBpedia [18] respectively, for experimental purpose. The datasets are constructed through the following steps:

(1) We first filter out all attribute triples, since such triples do not represent the relations of entities or concepts. After randomly sample some relational triples from the rest of the filtered dataset since original YAGO and DBpedia both have large collections of instance-view triples.
(2) After we obtain the entity set of instance view, we extract cross-view alignment of those entities to the ontology view of the two KBs. As a result, a portion of entities are linked to the associated concepts, which are naturally the nodes in the ontology view.
(3) Given all the associated concepts from step (2), we construct the corresponding ontology views base on the intersecting subgraph of the original ontologies.

It is noteworthy that the original YAGO has a taxonomical ontology with only three types of semantic relations, which casts limitation on semantic relations among concepts. Therefore, we enrich the ontology view of YAGO using the knowledge from ConceptNet [34], another KB which contains a large collection of meta-relations among concepts. The concepts in ConceptNet and YAGO are easily aligned by the shared WordNet-based IDs or concept names. Consequently, we obtain two datasets that are much larger than FB15K – the widely adopted instance KG benchmark dataset by many recent works [2, 19, 25, 42].



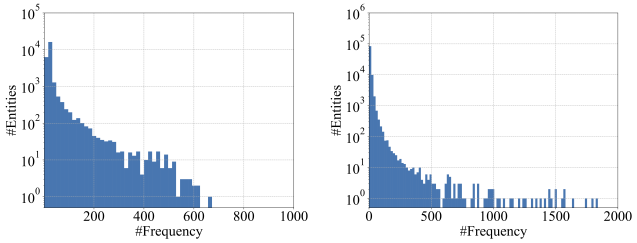
(a) JOIE-TransE-CT (With negative sampling)



(b) JOIE-TransE-CT (Without negative sampling)

Figure 7: Visualize effects on embeddings of negative sampling on cross-view links

As stated in Section 4.4, the frequency of entities and relations often follow a long-tail distribution (Zipf's law) in both YAGO26K-906 and DB111K-174 datasets, which is confirmed by the histogram in Figure 8.



(a) Histogram: YAGO26K-906 Entity Frequency

(b) Histogram: DB111K-174 Entity Frequency

Figure 8: Long-tail distribution holds on entity frequency from both YAGO26K-906(a) and DB111K-174(b)