Knowledge Base Embedding By Cooperative Knowledge Distillation

Raphaël Sourty, Jose G. Moreno, Francois-Paul Servant, Lynda Tamine

Knowledge Base Embedding By Cooperative Knowledge Distillation

Raphael Sourty[1,2], **Jose G. Moreno**[1], **Francois-Paul Servant**[2], **Lynda Tamine**[1]
[1]Université Paul Sabatier, IRIT, Toulouse, France
[2]Renault
{raphael.sourty, jose.moreno, tamine}@irit.fr
francois-paul.servant@renault.com

Abstract

Knowledge bases are increasingly exploited as gold standard data sources which benefit various knowledge-driven NLP tasks. In this paper, we explore a new research direction to perform knowledge base (KB) representation learning grounded with the recent theoretical framework of knowledge distillation over neural networks. Given a set of KBs, our proposed approach KD-MKB, learns KB embeddings by mutually and jointly distilling knowledge within a dynamic teacher-student setting. Experimental results on two standard datasets show that knowledge distillation between KBs through entity and relation inference is actually observed. We also show that cooperative learning significantly outperforms the two proposed baselines, namely traditional and sequential distillation.

1 Introduction

Knowledge Bases (KB), organizing structured information about entities (nodes), and relations (edges) as graphs, are increasingly exploited as gold standard data sources for a broad range of human-AI tasks including language modeling (Logan et al., 2019), question answering (Shen et al., 2019) and semantic search (Bast et al., 2016). Although typical KBs may include a huge amount of observed knowledge through millions of entities and their relations, they are by nature incomplete since they can only capture a fraction of world knowledge. This limitation has given rise to extensive research work that focuses on the issue of predicting new knowledge from the observed one (Socher et al., 2013; Nickel et al., 2016). This issue has been successfully tackled by neural approaches for representation learning of KBs (Wang et al., 2017; Sun et al., 2019; Bordes et al., 2013; Yang et al., 2014). These models aim at representing KB entities and relations in low-dimensional embedding spaces, and supporting relational inferences using simple vector algebra. Recent years have witnessed increasing interest toward embedding models leveraged to connect multiple KBs (Liu et al., 2016; Chen et al., 2017; Trivedi et al., 2018; Zhu et al., 2017; Zhang et al., 2019).

The key objective of multi-graph representation learning is to empower the entity and relation models with different graph contexts that potentially bridge different semantic contexts. To achieve this goal, embeddings are learned upon the combined triples across graphs. Although the above multi-graph representation learning methods have achieved promising results, they are still challenged by two main limitations. First, they are particularly suited to graph-alignment and machine translation as downstream tasks which necessarily leads to tackle computational challenges in large-scale KBs. Second, these methods assume that each KB has access to all the entities and relations that are stored in the other KBs, while it may not be feasible neither relevant to the KBs to share unaligned information such as in personal KBs (Balog and Kenter, 2019).

Following a different objective, we argue that apart from any downstream task, modeling the relational patterns across KBs might mainly focus on explicitly modeling connectivity patterns within each KB using its own observed triples and, infer additional patterns from its peer by only using partially aligned observed triples. As a motivating example let us consider two KBs (KB^1

Proceedings of the 28th International Conference on Computational Linguistics, pages 5579–5590
Barcelona, Spain (Online), December 8-13, 2020

and KB^2) that contain facts regarding cities, capitals, and countries as illustrated in Figure 1, but where none of them include the fact that *Rome is a city of Italy*. By learning an embedded space, KB^1 model may be able to correctly generalize the relation *CityIn* and infer that *Rome is a city of Italy* by grouping other similar entity embeddings to *Rome* such as the embedding of *Pisa*. Although this knowledge was not inferred by KB^1, KB^2 can teach this information to KB^1 by distillation. Then, KB^1 model will be able to understand the relation *CapitalOf* by directly observing examples within its own semantic context and the relation *CityIn* by distilled knowledge from KB^2 semantic context.

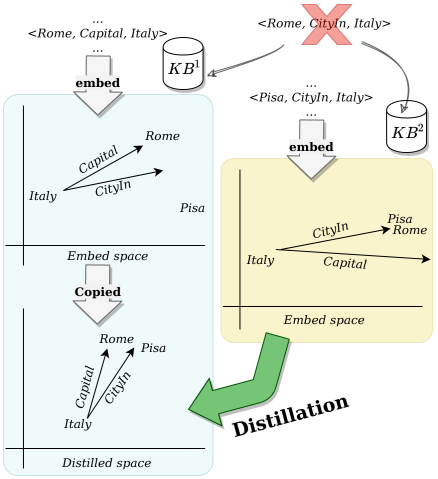

Figure 1: Motivating example.

Accordingly, unlike existing work in multi-graph embedding which relies on a unified view over multiple graphs, our work rather relies on multiple within-KB views that are bridged with aligned information. While each KB may be learning embeddings on its own semantic context based on associated hard triples, it can additionally exchange inferred knowledge from soft aligned triples provided by other KBs, in turn improving the embeddings of each other based on different semantic contexts. Our key idea is to model a knowledge distillation process (Hinton et al., 2015) across KBs to empower their generalization ability. Despite the number of works that show the rationale behind the entity and relation inference between KBs (Sun et al., 2018; Zhu et al., 2017), none has shown the feasibility of the knowledge distillation framework to model knowledge inference between KBs. Since the KBs play symmetric roles in knowledge transfer, one critical issue is how to train each KB model using entity/relation labels based on soft predictive distributions provided by the teacher, as well as its own predictive distribution. To tackle this issue, we argue for a mutual learning paradigm (Zhang et al., 2018), where each KB acts dynamically as either a teacher or a student. Unlike traditional static one-way knowledge transfer from a teacher model to a student model, we argue towards a two-way cooperative knowledge transfer between a KB and its peers. Concretely our set up is the following: the representation learning model of each KB is equipped with two losses which are jointly optimized: 1) a classic KB supervised margin based ranking loss whose objective is to make the scores of positive triples lower than those of negative ones; and 2) a mimicry cooperative distillation loss that makes the posterior class predictions of aligned entities and aligned relations close to the entity and relation class probabilities of its peer respectively. Through joint optimization, the knowledge is also naturally transferred from the seed aligned information to unaligned information.

In summary, the main contributions of the paper are the following: 1) a first attempt to ground multi-graph representation learning by a knowledge distillation theoretical framework; 2) a novel KB representation learning model called KD-MKB, based on a cooperative knowledge distillation strategy; 3) experiments on two standard datasets, WN18RR and FB15K-237, that empirically validate the rationale of knowledge distillation across KBs and show the effectiveness of the cooperative knowledge distillation as proposed in *KD-MKB*.

The remainder of this paper is structured as follows. Section 2 presents the related works. In Section 3, we first introduce the used preliminary notions and then detail the *KD-MKB* model. In Section 4, we present and discuss the experimental results. Section 5 concludes the paper.

2 Related Work

2.1 Representation learning across multiple KBs

Learning KB embeddings has drawn a huge attention in recent years. The embedding models are mainly categorized into: 1) translational models such as TransE (Bordes et al., 2013), TransH (Wang et al., 2014), and TransR (Lin et al., 2015) or complex based models such as ComplEx (Trouillon et al., 2016) and RotatE (Sun et al., 2019), which learn vector embeddings of both entities and relations by interpreting

a relation as a translation operation from a head entity to a tail; 2) deep neural models based for instance on graph convolutional networks (Schlichtkrull et al., 2018). More recently, representation learning across multiple graphs has gained an increasing attention. The general objective is to encode different graphs into a unified embedding space, such that the alignment likelihood between entities can be directly measured via their embeddings. Trivedi et al. (Trivedi et al., 2018) propose the *LinkNBed* multi-graph representation learning model based on a deep neural architecture. *LinkNBed* jointly learns relational data embeddings across multiple graphs in a shared space and entity linkage between these graphs using a multi-tasking approach. Sun et al. (Sun et al., 2018) propose a bootstrapping approach for entity alignment across multiple graphs. The key idea is to iteratively label likely alignments as training data and use them to further improve entity embeddings and alignment. Other work extends embedding models to multilingual learning across graphs. A seminal work is *MTransE* (Chen et al., 2017) which connects monolingual models by jointly aligning cross-lingual counterparts. On the other hand, in (Chen et al., 2018), the authors propose a co-training process combining multiple multilingual graph embedding models to learn on two views namely the structure and literal descriptions of entities. None of these methods jointly learn multiple KB embeddings while preserving the structure of each KB, which is the main goal of our work.

2.2 Knowledge Distillation

Knowledge distillation has been initially designed to distill the function approximated by a powerful ensemble of models playing the role of teacher, to a simpler single model playing the role of student (Bucila et al., 2006). This idea has given recently rise to an increasing attention for distilling the generalization ability from a large and easy-to train network model to a small but harder to train network (Adriana et al., 2015). The general framework relies on training a teacher first and then uses a teacher outputs in the form of posterior class probabilities to train the student model such as it mimics the teacher by providing similar outputs. Knowledge distillation has been widely used in NLP tasks to distill large models into small models (Mou et al., 2016) or ensembles of models into single models (Liu Yijia, 2018; Liu Xiaodong, 2019; Kevin Clark, 2019). Mou et al. (Mou et al., 2016) addressed the problem of distilling word embeddings in NLP applications. They proposed a supervised encoding approach to distill task-specific knowledge from cumbersome word embeddings. The approach has been shown to be effective in sentiment analysis and relation classification. Clark et. al (Kevin Clark, 2019) rather distill knowledge from single-task teacher models to multi-task student models. Their work extends born-again networks (Tommaso Furlanello, 2018) to the multi-task setting. The authors mainly rely on the teacher annealing technique, which consists in mixing the teacher prediction with the ground truth label during training. This strategy allows the student surpassing the teacher. The method has shown good performance in various NLP tasks including textual entailment, question-answering and paraphrase. In all of these works, distillation is applied on a pair of models that statically play either the role of teacher or student. In contrast, we adopt a mutual learning approach proposed in computer vision (Zhang et al., 2018) where a set of models dynamically play the role of teacher-student. However, unlike the learning framework proposed in (Zhang et al., 2018), we rather propose a cooperative learning over different learning tasks with associated respective ground truths and where knowledge distillation enables communication between models via shared data characteristics. Beyond, to the best of our knowledge, it is the first work that models and empirically validates the concept of knowledge distillation across KBs.

3 Cooperative Knowledge Distillation Across Multiple KBs

In this section, we describe *KD-MKB*, a KB representation learning model. We first introduce a couple of terminological definitions so we can formally define our model.

Knowledge base. A knowledge base KB represents a graph $(\mathcal{E}, \mathcal{R})$ which includes a set of entities $\mathcal{E} = \{e_1, e_2, \dots e_{N_e}\}$, a set of relations $\mathcal{R} = \{r_1, r_2, \dots r_{N_r}\}$, and a set of real relation facts as positive triples (e_x, r_w, e_y) denoted T^+ among all the possible ones in $\mathcal{E} \times \mathcal{R} \times \mathcal{E}$. The set of negative triples is denoted T^-.

Aligned entities and relations. Let $\mathcal{KB} = \{KB^1, KB^2, \dots KB^n\}$ represent a collection of KBs. For

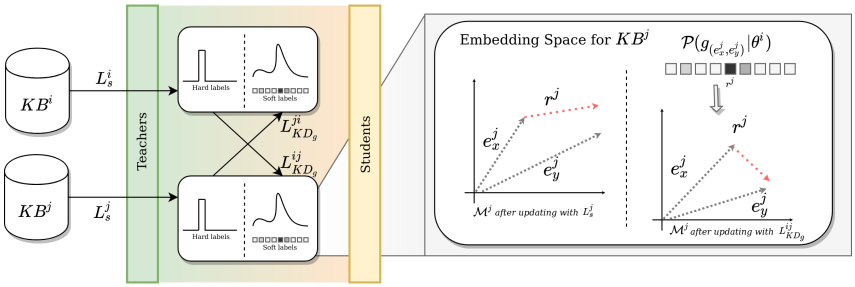

Figure 2: KD-MKB model architecture. The zoom in over the $\mathcal{M}^j$ model is illustrated with a relation distillation example.

a pair $(KB^i, KB^j) \in \mathcal{KB}^2$, KB^i (resp. KB^j) comprises a set of $\mathcal{E}^i$ (resp. $\mathcal{E}^j$) entities, a set of $\mathcal{R}^i$ (resp. $\mathcal{R}^j$) relations. $I_e(i,j) = \{(e_x^i, e_y^j) \in \mathcal{E}^i \times \mathcal{E}^j\}$ is the set of aligned entities meaning that e_x^i and e_y^j represent the same real world entity. Note that the set of entities from KB^i, denoted $I_e^i(i,j)$, is equal to its counterpart $I_e^j(i,j)$, and $|I_e(i,j)| = |I_e^i(i,j)|$. Similarly, $I_r(i,j) = \{(r_v^i, r_w^j) \in \mathcal{R}^i \times \mathcal{R}^j\}$ denotes the set of aligned relations such that r_v^i and r_w^j represent equivalent relations, and $|I_r(i,j)| = |I_r^i(i,j)|$.

3.1 Knowledge distillation from a KB to its peer

We conjecture that knowledge transfer between KBs can be drawn by relation and entity distillation. Our underlying intuitions are the following:

Relation distillation. Let us consider (e_1^i, e_x^j), $(e_2^i, e_y^j) \in I_e(i,j)$ two pairs of aligned entities between KB^i and KB^j. Assuming the existence of aligned relations between KB^i and KB^j, our intuition is that such entity pairs lead to the same probability of relation inference because the aligned entities refer to the same real world objects (Sun et al., 2018; Zhu et al., 2017). Accordingly, we argue for the relevance of mutually distilling likely aligned relations from one KB to its peers. Formally, plausibility scores of triples (e_1^i, r^i, e_2^i) can be estimated with high confidence based on plausibility scores of triplets (e_x^j, r^j, e_y^j) and vice versa.

Entity distillation. Let us consider $(r_v^i, r_w^j) \in I_r(i,j)$ and $(e_1^i, e_x^j) \in I_e(i,j)$ a pair of aligned entities and relations between KB^i and KB^j. Similarly to the intuition underlying relation distillation, we believe that such relation pairs lead to the same probability of entity inference because the aligned relations bring equivalent semantics that link between entities (Sun et al., 2018; Zhu et al., 2017). Thus, we argue for the relevance of mutually distilling likely aligned entities from one KB to its peer. Analogously to relation distillation principle, plausibility scores of triples (e_1^i, r_v^i, e_2^i) can be estimated with high confidence based on plausibility scores of triples (e_x^j, r_w^j, e_y^j) and vice versa.

3.2 Formulation of the KD-MKB model

In this paper, we study the representation learning of entities and relations across multiple KBs, while preserving the essential information included in each KB. Formally, given a collection of KBs $\mathcal{KB} = \{KB^1, KB^2, \ldots KB^n\}$, a knowledge embedding model $\mathcal{M}^i$ is learned to preserve entities and relations of each $KB^i, i = 1 \ldots n$ in a separated embedding space.

3.2.1 Design principles and objectives

Our main design principle is to, on one hand, learn embeddings directly from knowledge included in each KB, and on the other hand, improving the learning using knowledge distilled from its peers w.r.t to aligned entities and aligned relations. Based on this principle, the learning framework jointly achieves two complementary objectives.

Objective O1. Preserve the relational structure of each KB. For each participating KB^i, a dedicated knowledge embedding model $\mathcal{M}^i$ takes triples (e_x^i, r^i, e_y^i) either positive in T_i^+ or negative triples in T_i^- and learns corresponding embedding vectors $(\mathbf{e_x^i}, \mathbf{r^i}, \mathbf{e_y^i})$ by maximizing a triple plausibility scoring function $f_i : \mathcal{E}^i \times \mathcal{R}^i \times \mathcal{E}^i$ in a k_i dimensional space. TransE (Bordes et al., 2013), TransH (Wang et al., 2014) and RotatE (Sun et al., 2019) are examples of state-of-the-art scoring functions.

Objective O2. Improve the generalization ability of the representation learning model of each KB by leveraging its peers. Based on a cooperative learning setting, each knowledge embedding model $\mathcal{M}^i$ is further improved using knowledge distilled from each of the other embedding models $\mathcal{M}^j, j = 1 \dots n, j \neq i$. Each KB model $\mathcal{M}^i$ acts dynamically as either a teacher or a student by respectively distilling or leveraging distilled relations and distilled entities from its peers.

Thus, we formulate the KD-MKB model with a set of n networks which act dynamically as either teacher or student networks and mutually learn each of them specific models $\mathcal{M}^i, i = 1 \dots n$. Figure 2 provides an overview of the KD-MKB architecture with a setting of 2 ($n = 2$) teacher-students.

Each KB model $\mathcal{M}^i$ uses a teacher-student setting which learns from the ground-truth labels using a score function that measures the plausibility of the embeddings and the soft-labels provided by the $n-1$ teacher networks as prediction outputs using the probability of relation inference and entity inference based on the principle of knowledge distillation. The probability mass associated with each prediction output provided by the other teacher KBs $KB^j, j = 1 \dots n, j \neq i$ allows the model $\mathcal{M}^i$ to learn richer contextual information about the relation and entity embeddings similarity, leading to an increased ability of generalization. Thus, the model $\mathcal{M}^i$ is equipped with two losses which are jointly optimized: a classic KB supervised loss L^i_s on ground-truth labels and a mimicry cooperative knowledge distillation loss L^i_{KD} on soft-labels.

$$\mathcal{L}(\theta^i) = (1-\alpha)L^i_s + \alpha L^i_{KD} \tag{1}$$

where α is a hyperparameter.

Accordingly, each KB model learns both to correctly predict the correct label based on ground-truth training triples (loss L^i_s) as well as to match the posterior probability estimate of relations and entities provided by its peers (loss L^i_{KD}), following the intuitions outlined above (see Section 3.1). Such mutual learning helps each KB to learn additional context from its peers.

3.2.2 Supervised classification loss

Following objective O1 (see Section 3.1), we adopt a standard KB embedding model, namely TransE (Bordes et al., 2013). It is worth mentioning that other KB embedding models can also be used (eg., TransH (Wang et al., 2014) or RotatE (Sun et al., 2019)). Given a relation fact (e^i_x, r^i_w, e^i_y) in KB^i, we use the following score function to estimate the plausibility of the embeddings:

$$f_i(\mathbf{e^i_x}, \mathbf{r^i_w}, \mathbf{e^i_y}) = - \parallel \mathbf{e^i_x} + \mathbf{r^i_w} - \mathbf{e^i_y} \parallel \tag{2}$$

where $\| \cdot \|$ denotes either L_1 or L_2 vector norm. Accordingly, we define, the probability of (e^i_x, r^i_w, e^i_y) being a true triple as follows:

$$\mathcal{P}(y_{(e^i_x, r^i_w, e^i_y)} = 1 \mid \theta^i) = sigmoid(f_i(\mathbf{e^i_x}, \mathbf{r^i_w}, \mathbf{e^i_x})) \tag{3}$$

where $y_{(e^i_x, r^i_w, e^i_y)}$ is a random variable with value 1 if triple (e^i_x, r^i_w, e^i_y) is true (ie. relation fact), and 0 otherwise, $sigmoid(f_i(\mathbf{e^i_x}, \mathbf{r^i_w}, \mathbf{e^i_x})) = \frac{1}{1+\exp(-f_i(\mathbf{e^i_x}, \mathbf{r^i_w}, \mathbf{e^i_x}))}$, the logistic sigmoid applied to each triple score. The embedding model parameters θ^i, are defined by minimizing the logistic loss function:

$$L^i_s = \sum_{(e_1, r, e_2) \in T^+_i \cup T^-_i} log(1 + \exp(-y_{(e_1,r,e_2)} f_i(\mathbf{e_1}, \mathbf{r}, \mathbf{e_2})) \tag{4}$$

3.2.3 Cooperative knowledge distillation loss

Following objective O2 (see Section 3.1), the knowledge distillation is cooperatively conducted on the set of n KBs. At each learning step, each KB model $\mathcal{M}^i$ takes turns in the student-teacher process. As a teacher, the model distills its knowledge background through class prediction estimates L^i_s which are used as soft labels by the other student KBs to compute their mimicry loss function $L^j_{KD}, j = 1 \dots n, j \neq i$. Mutually, as a student, $\mathcal{M}^i$ model uses in its own mimicry loss L^i_{KD} soft labels distilled from the other KB teachers through $L^j_s, j = 1 \dots n, j \neq i$. From the perspective of KB^i, the distillation loss function

L^i_{KD} is formalized as the sum of two losses related to relation distillation $L^{ij}_{KD_r}$ and entity distillation $L^{ij}_{KD_e}$ from teacher network j to student network i as follows:

$$L^i_{KD} = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} L^{ij}_{KD_r} + L^{ij}_{KD_e} \tag{5}$$

Following the relation (resp. entity) distillation principles, the distillation function $L^{ij}_{KD_r}$ (resp. $L^{ij}_{KD_e}$) quantifies the match of each student network relation (resp. entity) prediction outputs using soft labels provided by the teacher networks with respect to the plausibility of the embeddings estimate given by the corresponding supervised classification functions f_j on ground truth labels. Relation and entity confidence scores used by the student and the teachers for prediction are obtained by converting the triple plausibility scores f_i and f_j on triples involving seed aligned relations and seed aligned entities as detailed in the following.

Relation distillation. A relation distillation favors the student model $\mathcal{M}^i$ to mimic the teacher model $\mathcal{M}^j$ on the relation prediction outputs over the set of aligned relations $r \in I_r(i,j)$ such as triples (e^j_1, r, e^j_2) and (e^i_1, r, e^i_2) have close plausibility scores. Thus, $L^{ij}_{KD_r}$ is computed as follows:

$$L^{ij}_{KD_r} = \sum_{(e^j_x, r, e^j_y) \in T^+_j : (e^i_x, e^j_x), (e^i_y, e^j_y) \in I_e(i,j)} \mathcal{D}(\mathcal{P}(r_{(e^j_x, \cdot, e^j_y)} \mid \theta^j), \mathcal{P}(r_{(e^i_x, \cdot, e^i_y)} \mid \theta^i)) \tag{6}$$

where $\mathcal{D}$ is the distillation function which can be defined in several ways (Sau and Balasubramanian, 2016) such as the L2 loss (Ba and Caruana, 2014) or Kullback-Leiber divergence (Hinton et al., 2015), $r_{(e_x, \cdot, e_y)}$ is a categorical variable with $|I_r(i,j)|$ values corresponding to aligned relation labels, $\mathcal{P}(r_{(e^j_x, \cdot, e^j_y)} \mid \theta^j)$ is a categorical distribution generated from the true triples $(e^j_x, r, e^j_y) \in T^+_j$ and $\mathcal{P}(r_{(e^i_x, \cdot, e^i_y)} \mid \theta^i)$, a categorical distribution generated from the triplets involving soft relation labels provides by model $\mathcal{M}^j$. The relation confidence score of relation r_v is obtained by converting plausibility scores using the softmax function over the aligned relations $r \in I_r(i,j)$ as below:

$$\mathcal{P}^v(r_{(e_x, r_v, e_y)} \mid \theta^k) = softmax(f_k(\mathbf{e^k_x}, \mathbf{r_v}, \mathbf{e^k_x})) \tag{7}$$

where $k = i, j$ and, $softmax(f_k(\mathbf{e^k_x}, \mathbf{r_v}, \mathbf{e^k_x})) = \frac{\exp(f_k(\mathbf{e^k_x}, \mathbf{r_v}, \mathbf{e^k_x}))}{\sum_{r_w \in I_r(i,j)} \exp(f_k(\mathbf{e^k_x}, \mathbf{r_w}, \mathbf{e^k_x}))}$, the softmax function applied to each triple score.

Entity distillation. An entity distillation favors the student model $\mathcal{M}^i$ to mimic the teacher model $\mathcal{M}^j$ on the link prediction outputs over the set of aligned entities $e \in I_e(i,j)$ such as triples (e^j_x, r^j_w, e^j_y) and (e^i_1, r^i_v, e^i_2) have close plausibility scores. Thus, $L^{ij}_{KD_e}$ is computed as follows:

$$L^{ij}_{KD_e} = \sum_{(e^j_x, r, e^j_y) \in T^+_j : (e^i_x, e^j_x) \in I_e(i,j), (r^i_v, r^j_v) \in I_r(i,j)} \mathcal{D}(\mathcal{P}(e_{(e^j_x, r^j_v, \cdot)} \mid \theta^j), \mathcal{P}(e_{(e^i_x, r^i_v, \cdot)} \mid \theta^i)) \tag{8}$$

$r_{(e_x, r, \cdot)}$ is a categorical variable with $|I_e(i,j)|$ values corresponding to aligned entity labels, $\mathcal{P}(r_{(e_x, r, \cdot)} \mid \theta^j)$ is a categorical distribution generated from the true triples $(e^j_x, r, e^j_y) \in T^+_j$ and $\mathcal{P}(r_{(e^j_x, r, \cdot)} \mid \theta^i)$, a categorical distribution generated from the triplets involving soft relation labels provides by models $\mathcal{M}^j$. The entity confidence score of the entity e_y is obtained by converting plausibility scores using the softmax function over the aligned entities $e \in I_e(i,j)$ as below:

$$\mathcal{P}^y(r_{(e_x, r, e_y)} \mid \theta^k) = softmax(f_k(\mathbf{e_x}, \mathbf{r}, \mathbf{e_y})) \tag{9}$$

where $softmax(f_k(\mathbf{e_x}, \mathbf{r}, \mathbf{e_y})) = \frac{\exp(f_k(\mathbf{e_x}, \mathbf{r}, \mathbf{e_y}))}{\sum_{e_z \in I_e(i,j)} \exp(f_k(\mathbf{e_x}, \mathbf{r}, \mathbf{e_z}))}$, the softmax function applied to each triple score.

Algorithm 1: KD-MKB training model

Input:
KD-MKB parameter α, KB embedding model with own parameters
Set of knowledge bases $\mathcal{KB} = \{KB^1, KB^2, \ldots, KB^n\}$
Set of aligned entities and relations $I_e(i,j)$, $I_r(i,j)$ $\forall i,j \in \{1,..,n\}$ with $i \neq j$

Initialize:
Set of models $\mathcal{M} = \{\mathcal{M}^1, \mathcal{M}^2, \ldots, \mathcal{M}^n\}$

while *convergence or maximum number of iterations not achieved* **do**
  for $\mathcal{M}^i \in \mathcal{M}$ **do**
    index_top-k ←create index with $\mathcal{M}^i$
    $batch^i \leftarrow$ sample triplets from T_i^+ and T_i^-
    $\mathcal{L}^i \leftarrow (1-\alpha) \times \mathcal{L}^i$ w.r.t $batch^i$ following eq. (4)
  for $\mathcal{M}^i \in \mathcal{M}$ **do**
    for $\mathcal{M}^j \in \mathcal{M}$ **do**
      if $i \neq j$ **then**
        $batch_e^j \leftarrow$query index_top-k using $batch^i \cap I_e(i,j)$ to get $\mathcal{P}(g_{(e_x^i, e_y^i)}|\theta^i)$
        $batch_r^j \leftarrow$query index_top-k using $batch^i \cap I_r(i,j)$ to get $\mathcal{P}(g_{(e_x^i, r_v^i)}|\theta^i)$
        $L_{KD}^j \leftarrow$ distill teacher $\mathcal{M}^i$ to student $\mathcal{M}^j$ with eq. (5) w.r.t. $batch_r^j$ and $batch_e^j$
        $\mathcal{L}^j \leftarrow \mathcal{L}^j + \alpha \times L_{KD}^j$
  for $\mathcal{M}^i \in \mathcal{M}$ **do**
    Jointly update $\mathcal{M}^i$ w.r.t. $\mathcal{L}^i$

Output: Parameters θ^i for each model $\mathcal{M}^i$

3.3 The training procedure

A key characteristic of our proposed cooperative knowledge distillation is that all the losses $\mathcal{L}(\theta^i), i{=}1 \ldots n$ of the n knowledge embedding models, are jointly and cooperatively optimized. At each iteration, each loss $\mathcal{L}(\theta^i)$ uses both the true labels and the soft labels provided by network models $\mathcal{M}^j, j{=}1 \ldots n, j \neq i$ to update parameters θ^i. The training model is summarized in Algorithm 1. The learning strategy is setup in each mini-batch based model update. At each iteration, all the losses $\mathcal{L}(\theta^i)$ are jointly learned using one mini-batch for training L_s^i and $(n-1)$ mini-batches comprising pairs of alignments for training $L_{KD_e}^{ij}$ and $L_{KD_r}^{ij}$. Since the sizes of entity and relation sets used in the softmax normalization calculation of $\mathcal{P}(.)$ over $\mathcal{E}^i$ or $\mathcal{R}^i$ may be very large, we apply a sampling technique to estimate the probability distribution as done in previous work (Liu Yijia, 2018). It consists in selecting the top k candidate entities (resp. relations) w.r.t. equation 2 to the given example to be distilled plus k random entities (resp. relations). The teacher is in charge of the choice of the top-k candidates. Thus, we only use $2 \times k$ entities (resp. relations) for the softmax normalization instead of $|\mathcal{E}^i|$ or $|I_e(i,j)|$ (resp. $|\mathcal{R}^i|$ or $|I_r(i,j)|$) total values which drastically reduce the number of required computations for each distillation mini-batch.

4 Experiments

Two main objectives have guided our experiments: 1) show the validity of knowledge distillation to formally support knowledge transfer between KBs; 2) evaluate the effectiveness of the KD-MKB model.

4.1 Settings

Datasets and splits. We perform our experiments on two standard real-world WN18RR and FB15K-237 KBs[1]. We simulate the multiple KBs setting by randomly splitting each of the WN18RR and FB15K-237 KB train triples into 2 and 3 partitions ($n = 2, 3$ in the KD-MKB setting, see Section 3.2). Our motivation behind this evaluation setting is sustained by two reasons: 1) our goal with *KD-MKB* is to learn empowered KB embeddings instead of multi-graph embeddings; 2) evaluate the intrinsic effect of the *KD-MKB* model without any bias induced by uncontrolled effect of knowledge alignment quality. More

[1]Both KBs are available at https://www.microsoft.com/en-us/download/details.aspx?id=52312 (Last checked 26/10/2020)

Dataset	# Entities	# Relations	# Train	# Val	# Test
WN18RR	40,923	11	86,834	3,033	3,134
FB15K-237	14,255	237	272,115	17,535	20,466

Table 1: Statistics of WN18RR and FB15K-237 datasets.

precisely, two FB15K-237 partitions usually share 95% of the entities but drops to 64% for WN18RR. The setting $n = 1$ allows reporting the traditional KB model on the entire set of triples. Table 1 provides statistics for both datasets used in our experiments. Within each setting, we obtain n teacher $\mathcal{M}^i$ and student $\mathcal{M}^j$ models, so reported results are averages. We compare the performance of our model using state-of-the art neural representation models, namely TransE (Bordes et al., 2013). We focus on the standard entity link prediction task for knowledge base population. This task evaluates the model performances for a given tail query (e_i,r_j,$e?$) where the response is a ranked list of entities that better fit $e?$ (similarly, head queries can be evaluated). We use the standard HITS@k ($k = 1, 3, 10$) and Mean Reciprocal Rank (MRR) metrics. We report the means of multiple runs over test partitions.

Knowledge distillation strategies. We analyze the effectiveness of knowledge distillation strategies by comparing the results reported on the following scenarios: 1) *Independent* is the traditional one-way distillation setup (Hinton et al., 2015) where only half ($n = 2$) or third ($n = 3$) of the knowledge is transferred from the teacher to the respective student. The teacher model is pre-trained and provides posterior entity and relation predictions to the student model. Note that within this setting, each KB model plays statically either the role of teacher or student during the learning process; 2) *Xdistills∼X* is a sequential setup in which first each model is trained over one of the set defined in the partition until convergence. Then he plays the role of a teacher and distills it's knowledge to other models that play the role of students, then, it plays the role of student. Thus the KB teacher and the KB student models parameters are updated one after the other in a sequential fashion[2]; 3) *KD-MKB* model in which each model dynamically and simultaneously acts as teacher and as student during the whole training process. Thus, the predictions and parameters of the KB models are jointly updated.

Implementation details. We implemented all baselines and our model using PyTorch[3]. The loss function is minimized using the Adam stochastic method with a learning rate of 10^{-5}. The maximum number of iterations is set to 8×10^4. Parameters of the TransE model were fixed by selecting the best configuration in the validation partition of each dataset and by following recommendations from (Sun et al., 2019). Thus, embedding size is fixed to 1000 (resp. 500), batch size to 512 (resp. 256), negative sampling size to 128 (resp. 512), the α^{al} adversarial loss to 1 (resp. 0.5), and the margin hyperparameter γ to 9 (resp. 6) for FB15K-237 (resp. WN18RR). Top-k entities are found using faiss (Johnson et al., 2017) with k set to 10. Hyperparameter α in Equation 1, is set to 0.98 following a loss analysis between L_s^i and L_{KD}^i[4].

4.2 Results and discussion

4.2.1 Does knowledge distillation between KBs work?

To the best of our knowledge, this is the first attempt in the literature to empirically assess about knowledge inference between KB embeddings using the theoretical framework of knowledge distillation (Hinton et al., 2015). Table 2 shows link prediction performances for the two used data sets when performing traditional independent distillation from a teacher $\mathcal{M}^i$ to a student $\mathcal{M}^j$. We can see from Table 2 that overall the performance levels of the student model follow those of the teacher model for all the metrics and that the performance trends remain the same for increasing numbers of KBs. This result empirically validates our idea about the modeling of knowledge inference between KBs through the formalization of simultaneous distillation of entities ($L_{KD_e}^{ij}$) and relations ($L_{KD_r}^{ij}$).

4.2.2 KD-MKB model analysis

Distillation model. To highlight the benefit behind cooperatively distilling knowledge across KBs, we report in Table 3 link prediction results using the three knowledge distillation strategies *Independent*,

[2]Other baselines can be proposed. However, using a sequential distillation guarantees that the baseline sees the triplets of one partition. The main drawback is that some of the knowledge learned as a teacher may be loss when behaving as a student.

[3]Our implementation is publicly available at https://github.com/raphaelsty/mkb

[4]Exploration of more elaborated loss combinations such as the one reported in (Zoph et al., 2020) is left for future work.

	WN18RR								FB15K-237							
	Teacher				Student				Teacher				Student			
n	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR	H@1	H@3	H@10	MRR
1	1.43	39.61	52.63	0.22	1.37	39.53	52.40	0.22	22.53	36.27	52.15	0.32	22.46	36.33	52.24	0.32
2	0.65	19.90	28.36	0.11	0.62	19.75	28.35	0.11	19.28	31.56	46.23	0.28	19.18	31.35	45.99	0.28
3	0.58	12.12	18.61	0.07	0.59	11.84	18.10	0.07	17.19	28.35	42.59	0.26	17.09	28.18	42.11	0.25

Table 2: Performance results for WN18RR and FB15K-237 datasets on the link prediction tasks using the traditional independent distillation model. H@k stands for HITS@k. n indicates the number of KB partitions used to learn the Teacher representation. For $n > 1$, reported values correspond to average performances of the multiple models over test sets.

Xdistills∼X and *KD-MKB* by using the same splits than those presented in Table 2. We report best and worst results of $\mathcal{M}^i$ and $\mathcal{M}^j$ models for each dataset partition. The main observation that can be drawn from Table 3 is that the *KD-MKB* model outperforms the *Independent* model (e.g., between 24.8% and 455.7% improvement based on HITS@1, between 17.9% and 85.7% improvement based on MRR) over all the partitions and datasets and w.r.t. to all the metrics. Additionally, we can also see that the *Xdistills∼X* model outperforms the *Independent* model. For example, when $n = 2$, the HITS@3 performance reaches a level around 27.93 (resp. 32.60) using *Xdistills∼X* for the WN18RR (resp. FB15K-237) dataset vs. a lower value around 19.95 (resp. 31.92) using the *Independent* model. This can be easily explained as the *Independent* model is only trained over the soft-labels in just one partition while the *Xdistills∼X* model uses first hard labels from its own partition (when it plays the role of a teacher) and then uses soft labels from the other partitions (when it plays the role of a student). We can also interestingly observe that the *KD-MKB* model outperforms the *Xdistills∼X* model though by a lower percent change (e.g., between 1.0% and 11.0% improvement for HITS@10 using WN18RR dataset, between 10.6% and 18.6% improvement for HITS@10 using FB15K-237 dataset). It is worth mentioning that the *KD-MKB* model uses the same number of soft- and hard-labels than the *Xdistills∼X* model but our proposed cooperative strategy takes more advantage of both kind of labels. This result highlights a clear benefit of both dynamically switching between the teacher and student roles for each of the $\mathcal{M}^i$ models empowered by the mimicry cooperative learning and joint update of their parameters.

Distillation strategy	n	HITS@1			HITS@3			HITS@10			MRR		
		Best	Worst	%Chg.	Best	Worst	%Chg	Best	Worst	%Chg	Best	Worst	%Chg
Independent	2	0.70	0.60	+455.7%	19.95	19.86	+68.6%	28.38	28.35	+48.6%	0.11	0.11	+72.7%
	3	0.74	0.28	+173.0%	12.53	11.59	+68.4%	19.78	17.32	+70.5%	0.07	0.07	+85.7%
Xdistills∼X	2	1.21	1.03	+221.5%	27.93	27.69	+20.4%	41.75	41.48	+1.0%	0.16	0.15	+18.8%
	3	1.57	1.26	+28.7%	16.51	16.19	+27.8%	30.47	30.42	+11.0%	0.10	0.10	+30.0%
KD-MKB	2	3.89	3.74	-	33.64	33.55	-	42.18	42.02	-	0.19	0.19	-
	3	2.02	1.75	-	21.10	19.57	-	33.72	32.72	-	0.13	0.12	-

(a) WN18RR dataset

Distillation strategy	n	HITS@1			HITS@3			HITS@10			MRR		
		Best	Worst	%Chg.	Best	Worst	%Chg	Best	Worst	%Chg	Best	Worst	%Chg
Independent	2	19.38	19.18	+24.8%	31.92	31.21	+18.3%	46.45	46.02	+14.7%	0.28	0.28	+17.9%
	3	17.32	17.03	+41.1%	28.55	28.22	+32.7%	42.88	42.43	+23.7%	0.25	0.25	+36.0%
Xdistills∼X	2	20.43	20.39	+18.4%	32.60	32.04	+15.8%	48.15	47.85	+10.6%	0.29	0.29	+13.8%
	3	18.85	18.70	+29.7%	29.61	29.50	+27.9%	44.73	44.46	+18.6%	0.27	0.27	+25.9%
KD-MKB	2	24.18	24.12	-	37.75	37.66	-	53.26	53.22	-	0.33	0.33	-
	3	24.44	24.36	-	37.88	37.82	-	53.06	52.95	-	0.34	0.33	-

(b) FB15K-237 dataset

Table 3: Performance results for WN18RR and FB15K-237 datasets on the link prediction task using variants of the distillation models. Reported values are the best and worst performances obtained within each n dataset split configuration. %Chg. denotes the effectiveness improvement of the KD-MKB model w.r.t concurrent distillation model strategies under consideration based on the best performing model.

Distillation with larger alignments. We further analyze the effect of the size of entity alignments on *KD-MKB* performance. To be coherent with the results presented in the previous experiments, we keep the same partitions. Figure 3 plots the performance variations w.r.t. to all the metrics using the

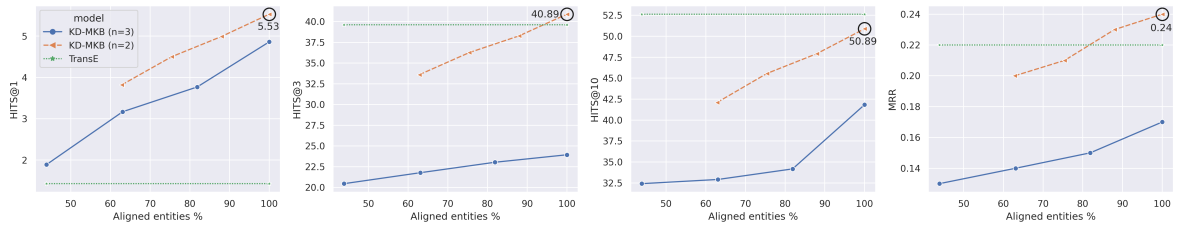

Figure 3: HITS@1, HITS@3, HITS@10 and MRR link prediction results for *KD-MKB* using WN18RR when different sizes of the alignment set ($I_e(i,j)$) are used. Our best performances are highlighted by a circle and values were included.

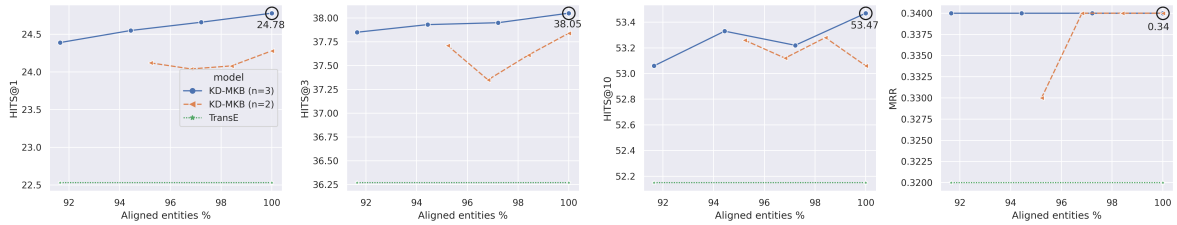

Figure 4: HITS@1, HITS@3, HITS@10 and MRR link prediction results for *KD-MKB* using FB15K-237 when different sizes of the alignment set ($I_e(i,j)$) are used. Our best performances are highlighted by a circle and values were included.

WN18RR dataset. It is worth mentioning, that unlike the FB15K-237 which exhibits an overlap of 95%, the WN18RR actually allows to simulate an increasing overlap of entities by adding extra information to $I_e(i,j)$ until reaching the 100% overlap. The additional mapping information between entities does not increase the number of triplets to train but it allows a larger number of entities to be distilled from teachers. Figure 3 indicates that in average, 20% of extra aligned entities leads to an absolute gain of 5.2 points in HITS@10 and 0.06 points in MRR. As expected, higher numbers of soft-labels improves the mutual knowledge inference from a KB to its peers. Results for FB15K-237 datasets are presented in Figure 4. In this case the increasing is beneficial being more stable when $n = 3$ because of the larger increase of the overlapped entities.

Multi-KB learning vs. single-KB learning. We depict in Figures 3 and 4 the performances of *KD-MKB* as well as for TransE using both datasets. Both models were trained using full train data, hard labels only for the latter, and hard-labels and soft-labels for the former. Results are constant for TransE as its performances do not depend on the number of aligned entities. When the number of aligned entities is 100% (highlighted by a circle), we can verify from these figures the improvements of using *KD-MKB* on multiple KBs instead of the classical TransE embedding model on an individual KB. For FB15K-237, both partitions configuration (e.g. *KD-MKB* with $n = 2$ and $n = 3$) outperform the TransE performances for all studied metrics. However, for WN18RR, *KD-MKB* slightly outperforms TransE when $n = 2$ in terms of HITS@3 and MRR, but fails to improve in terms of HITS@10.

5 Conclusion and Future Work

This paper presents a new framework for learning entity and relation embeddings over multiple KBs. Our framework exploits a new way to transfer learning from one KB model to its peers. First, we formalize entity and relation inference between KBs as a distillation loss over posterior probability distributions on aligned knowledge. Grounded on this finding, we propose and formalize a cooperative distillation framework where a set of KB models are jointly learned by using each of them hard labels from their own context and also soft labels provided by peers. We empirically demonstrate the rationale behind knowledge distillation between KBs and show the effectiveness of our cooperative learning framework on the link prediction task compared to the existing distillation strategies. Further experiments are planned for future work using more complex and realistic configurations for multiple KBs learning to assess about the generalizability of our findings. We also plan to extend our approach to consider weak alignments while distilling knowledge over pairs of KBs. This would empower the cooperative learning with higher generalization ability particularly for heterogeneous KBs. We believe that this work could be used without major change in the core methodology to support a wide range of knowledge-driven applications.

References

Romero Adriana, Ballas Nicolas, K Samira Ebrahimi, Chassang Antoine, Gatta Carlo, and B Yoshua. 2015. Fitnets: Hints for thin deep nets. *In Proceedings of International Conference on Learning Representations*.

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2654–2662. Curran Associates, Inc.

Krisztian Balog and Tom Kenter. 2019. Personal knowledge graphs: A research agenda. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 217–220, New York, NY, USA. Association for Computing Machinery.

Hannah Bast, Buchhold Björn, and Elmar Haussmann. 2016. Semantic search on text and knowledge bases. *Found. Trends Inf. Retr.*, 10(2–3):119–271, June.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2787–2795. Curran Associates, Inc.

Cristian Bucila, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, page 535–541, New York, NY, USA. Association for Computing Machinery.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 1511–1517. AAAI Press.

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. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 3998–4004. AAAI Press.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*.

Urvashi Khandelwal Christopher D. Manning Quoc V. Le Kevin Clark, Minh-Thang Luong. 2019. Bam! born-again multi-task networks for natural language understanding. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5931–5937.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2181–2187. AAAI Press.

Quan Liu, Hui Jiang, Zhen-Hua Ling, Si Wei, and Yu Hu. 2016. Probabilistic reasoning via deep learning: Neural association models. *CoRR*, abs/1603.07704.

Chen Weizhu Gao Jianfeng Liu Xiaodong, He Pengcheng. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496.

Zhao Huaipeng Qin Bing Liu Ting Liu Yijia, Che Wanxiang. 2018. Distilling knowledge for search-based structured prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 1393–1402.

Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. 2019. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy, July. Association for Computational Linguistics.

Lili Mou, Ran Jia, Yan Xu, Ge Li, Lu Zhang, and Zhi Jin. 2016. Distilling word embeddings: An encoding approach. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 1977–1980.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

Bharat Bhusan Sau and Vineeth N. Balasubramanian. 2016. Deep model compression: Distilling knowledge from noisy teachers. *CoRR*, abs/1610.09650.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, pages 593–607.

Tao Shen, Xiubo Geng, Tao Qin, Daya Guo, Duyu Tang, Nan Duan, Guodong Long, and Daxin Jiang. 2019. Multi-task learning for conversational question answering over a large-scale knowledge base. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2442–2451, Hong Kong, China, November. Association for Computational Linguistics.

Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pages 926–934.

Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping entity alignment with knowledge graph embedding. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4396–4402. International Joint Conferences on Artificial Intelligence Organization, 7.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*.

Michael Tschannen Laurent Itti Anima Anandkumar Tommaso Furlanello, Zachary Chase Lipton. 2018. Born-again neural networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1602–1611.

Rakshit Trivedi, Bunyamin Sisman, Xin Luna Dong, Christos Faloutsos, Jun Ma, and Hongyuan Zha. 2018. LinkNBed: Multi-graph representation learning with entity linkage. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 252–262, Melbourne, Australia, July. Association for Computational Linguistics.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. International Conference on Machine Learning (ICML).

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, page 1112–1119. AAAI Press.

Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge Graph Embedding: A Survey of Approaches and Applications. *IEEE Transactions on Knowledge and Data Engineering*, 29(12):2724–2743, December.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu. 2018. Deep mutual learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4328.

Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Multi-view knowledge graph embedding for entity alignment. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, pages 5429–5435, 08.

Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Iterative entity alignment via joint knowledge embeddings. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4258–4264. AAAI Press.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. 2020. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882*.