Co-training Embeddings of Knowledge Graphs and Entity Descriptions for Cross-lingual Entity Alignment

Muhao Chen[1], **Yingtao Tian**[2], **Kai-Wei Chang**[1], **Steven Skiena**[2] and **Carlo Zaniolo**[1]

[1]Department of Computer Science, University of California, Los Angeles
[2]Department of Computer Science, Stony Brook University
{muhaochen, kw2c, zaniolo}@cs.ucla.edu; {yittian, skiena}@cs.stonybrook.edu

Abstract

Multilingual knowledge graph (KG) embeddings provide latent semantic representations of entities and structured knowledge with cross-lingual inferences, which benefit various knowledge-driven cross-lingual NLP tasks. However, precisely learning such cross-lingual inferences is usually hindered by the low coverage of entity alignment in many KGs. Since many multilingual KGs also provide literal descriptions of entities, in this paper, we introduce an embedding-based approach which leverages a weakly aligned multilingual KG for semi-supervised cross-lingual learning using entity descriptions. Our approach performs co-training of two embedding models, i.e. a multilingual KG embedding model and a multilingual literal description embedding model. The models are trained on a large Wikipedia-based trilingual dataset where most entity alignment is unknown to training. Experimental results show that the performance of the proposed approach on the entity alignment task improves at each iteration of co-training, and eventually reaches a stage at which it significantly surpasses previous approaches. We also show that our approach has promising abilities for zero-shot entity alignment, and cross-lingual KG completion.

1 Introduction

Multilingual knowledge bases (KBs) such as DBpedia [Lehmann *et al.*, 2015], ConceptNet [Speer *et al.*, 2017], and Yago [Mahdisoltani *et al.*, 2015] constitute crucial sources of knowledge for AI-related applications. These KBs store knowledge graphs (KGs) that represent two aspects of structured knowledge: (1) the *monolingual knowledge* that models relational facts of entities as triples, (2) and the *cross-lingual knowledge* that synchronizes monolingual knowledge among multiple human languages (see Fig. 1). In addition to those, many KGs also store literal descriptions of entities in different languages [Xie *et al.*, 2016; Lehmann *et al.*, 2015].

Embedding models for KGs have been extensively studied in the past few years. These models aim at characterizing entities in low-dimensional embedding spaces, and supporting relational inferences for entity embeddings via simple vector
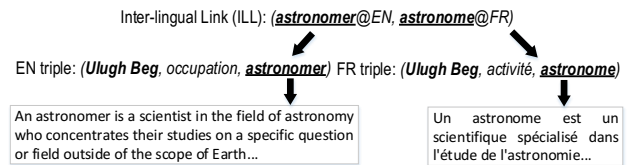

Figure 1: A simple example which shows triples, an ILL, and entity descriptions in a multilingual KG (DBpedia). The French description for *astronome* means *an astronomer is a scientist specialized in the study of astronomy*, which contains much fewer content details than the English description for *astronomer*.

algebra. Hence, they provide efficient and versatile methods to incorporate the symbolic knowledge of KGs into machine learning. Models of this kind have been widely applied to NLP-related tasks, such as relation extraction [Wang *et al.*, 2014], ontology population [Chen *et al.*, 2018], question answering (QA) [Bordes *et al.*, 2014], dialogue agents [He *et al.*, 2017], and visual semantic labeling [Fang *et al.*, 2017].

Recently, embedding models are leveraged to connect KG structures of multiple languages [Chen *et al.*, 2017a; 2017b; Sun *et al.*, 2017; Zhu *et al.*, 2017]. Emerging of such approaches is significant, inasmuch as they extend the inferences of KG embeddings to a multilingual scenario, and seek to benefit cross-lingual NLP tasks such as knowledge alignment, cross-lingual QA and machine translation. While such embeddings are generic and beneficial, it remains very challenging for corresponding approaches to precisely capture the cross-lingual inferences. The challenge is that the cross-lingual knowledge, which is typically formed as *inter-lingual links* (ILLs) that match cross-lingual counterparts of entities, is usually far from complete. In fact, ILLs cover less than 20% of the entities even in the most successful Wikipedia-based KBs. Hence, the lack of supervision by cross-lingual knowledge easily hinders the quality of cross-lingual inferences, which affects even more significantly when each language version of KG scales up and becomes inconsistent in contents and density.

While existing embedding models solely rely on the structured knowledge for cross-lingual learning, it would be promising to enhance the corresponding learning process with the literal descriptions of entities that are stored in many KGs [Xie *et al.*, 2016; Lehmann *et al.*, 2015; Mahdisoltani *et*

al., 2015]. These descriptions comprise an alternative view of entities that potentially bridges two languages, since the descriptions of an entity in different languages often share a lot of semantic information. However, it is non-trivial to characterize and utilize such information for cross-lingual learning, as this requires the model to learn to match descriptions across different languages with inadequate labels, while conquering the inconsistency of literals in content details, grammars, and word orders (as shown in Fig. 1). Moreover, aggregating semantic relatedness of descriptions from words of different languages is another challenge.

To address these issues, we propose a novel co-training-based approach KDCoE to enhance the semi-supervised learning of multilingual KG embeddings. KDCoE iteratively trains two component embedding models on multilingual KG structures and entity descriptions respectively. A KG embedding model jointly trains a translational knowledge model with a linear-transformation-based alignment model to encode the KG structure. A description embedding model employs an attentive gated recurrent unit encoder (AGRU) and multilingual word embeddings to characterize multilingual entity descriptions, and is trained to collocate the embeddings of cross-lingual counterparts. The co-training is processed on a large Wikipedia-based trilingual KG, for which a very small portion of ILLs is used for training. During each iteration of co-training, both models alternately propose a set of most confident new ILLs to strengthen the supervision of cross-lingual learning, which leads to gradually improved accuracy on cross-lingual inferences. Experimental results on entity alignment confirms the effectiveness of KDCoE that significantly outperforms previous models, while those results on zero-shot alignment and cross-lingual KG completion also show wider usability of our approach.

The rest of the paper is organized as follows. We first discuss the related work, and then introduce our approach in the section that follows. After that we present the experimental results, and conclude the paper in the last section.

2 Related Work

We discuss three lines of works that are relevant to this paper.
Monolingual KG Embeddings. KG embeddings are first explored in the monolingual scenario. The past half decade has seen much popularity on translational models, which mostly follow the forerunner TransE [Bordes *et al.*, 2013] to capture a triple (h, r, t) as a translation $\mathbf{r}$ between two entity embeddings $\mathbf{h}$, $\mathbf{t}$. Later works such as TransH [Wang *et al.*, 2014], TransR [Lin *et al.*, 2015], TransD [Ji *et al.*, 2015], and TransA [Jia *et al.*, 2016] differentiate such translations in separated spaces using different forms of relation-specific projections. Models of this family preserve well the KG structures in the embedding spaces regardless of their simplicity, and offer promising performance on KG completion and relation extraction tasks. In [Xie *et al.*, 2016] TransE is trained jointly with a convolutional neural network (CNN) to predict entities based on their descriptions. In addition to them, recent works also introduce successful non-translational models, such as DistMult [Yang *et al.*, 2015] and HolE [Nickel *et al.*, 2016] that adopt dot product and circular correlation

respectively, and neural models such as ConvE [Dettmers *et al.*, 2018]. These models perform comparably to or even better than translational models at the cost of model complexity.
Multilingual KG Embeddings. More recent work extends embedding models to multilingual learning on KGs. One representative work is MTransE [Chen *et al.*, 2017a]. MTransE connects monolingual models with a jointly trained alignment model, for which three aligment techniques are employed, i.e., axis calibration that adjusts embedding spaces to collocate cross-lingual counterparts (MTransE-AC), vector translation (MTransE-TV), and linear transformations across embedding spaces (MTransE-LT) for different languages. MTransE-LT thereof achieves the best performance on knowledge alignment tasks. JAPE is introduced in [Sun *et al.*, 2017] to strengthen the cross-lingual learning of MTransE-AC based on the similarity of entity attributes. This model performs well on KBs that provide numerical entity attributes, though such attributes are not generally available in many KBs. Another relevant model ITransE [Zhu *et al.*, 2017] incorporates self-training into a hard-alignment version of MTransE-AC. ITransE is used to align entities across monolingual KGs with coherent vocabularies and triples, but we find it does not adapt well to the inconsistent multilingual scenario. Note that off-line multilingual word embedding models, including LM [Mikolov *et al.*, 2013a], CCA [Faruqui and Dyer, 2014], and orthogonal-transformation-based OT [Xing *et al.*, 2015] can also be extended to KGs, but are outperformed by MTransE on cross-lingual tasks.
Co-training. Co-training combines multiple models to learn on different views of the data in the training process, in which all participating models take turn in suggesting more labels on unlabeled data to enhance the supervision. This technique is widely used in semi-supervised learning tasks, such as sentiment classification on bilingual corpora with incomplete labels [Wan, 2009], collaborative filtering in recommender systems with multiple user views [Zhang *et al.*, 2014], and semantic role labeling based on the semantic and syntactic views of documents [Thi *et al.*, 2016]. Our work conducts co-training on two views of the multilingual KG, i.e. structures and literal descriptions, which to the best of our knowledge, is the first work that incorporates co-training into embedding learning, as well as knowledge alignment tasks.

3 Modeling

We first provide the definition of multilingual KGs. In a KB, $\mathcal{L}$ denotes the set of languages, and $\mathcal{L}^2$ unordered language pairs. G_L is the language-specific KG of each language $L \in \mathcal{L}$. E_L and R_L respectively denote the corresponding vocabularies of entities and relations. $T = (h, r, t)$ denotes a triple in G_L such that $h, t \in E_L$ and $r \in R_L$. Boldfaced $\mathbf{h}$, $\mathbf{r}$, $\mathbf{t}$ represent the embedding vectors of head h, relation r, and tail t respectively. For a language pair $(L_1, L_2) \in \mathcal{L}^2$, $I(L_1, L_2)$ denotes a set of ILLs that align entities between L_1 and L_2, such that $e_1 \in E_{L_1}$ and $e_2 \in E_{L_2}$ for each ILL $(e_1, e_2) \in I(L_1, L_2)$. We assume the entity pairs have a 1-to-1 mapping and it is specified in $I(L_1, L_2)$. This assumption is congruent to the design of mainstream KGs [Lehmann *et al.*, 2015]. Besides the above structured knowledge, we

use D_L to denote the literal descriptions of entities in language L. A description $d_e \in D_L$ describes an entity $e \in E_L$ with a sequence of words from the word vocabulary W_L, i.e. $d_e = \{w_1, w_2, ..., w_l\}$.

KDCoE conducts iterative co-training of two components, i.e. the multilingual KG embedding model (KGEM) and the multilingual description embedding model (DEM), which capture embeddings with cross-lingual inferences for structured knowledge and entity descriptions respectively. During co-training, both components are trained in turns to propose new ILLs with high confidence, which populate the training set and become visible to future turns of training. We define the model on a pair of languages from $\mathcal{L}^2$ for which the ILLs are provided. For a KB with more than two languages, multiple models that bridge different languages compose the solution *w.l.o.g.* In the following subsections, we use a language pair $(L_i, L_j) \in \mathcal{L}^2$ to describe the definition of the model components and the entire learning process.

3.1 Multilingual KG Embeddings

The KGEM consists of two components that learn on the two facets of structured knowledge.

A knowledge model is learnt to preserve entities and relations of each language in a separated embedding space. Specifically, for each participating language L, a dedicated k_1-dimensional embedding space $\mathbb{R}_L^{k_1}$ is assigned for vectors of E_L and R_L. Like previous work [Chen *et al.*, 2017a; Sun *et al.*, 2017; Zhu *et al.*, 2017], we adopt the basic translational method of TransE for each involved language, which benefits the cross-lingual tasks with uniform representations of entities in different contexts of relations. The corresponding objective function is given as the following hinge loss,

$$S_K = \sum_{L\in\{L_i,L_j\}} \sum_{(h,r,t)\in G_L \wedge (\hat{h},r,\hat{t})\notin G_L} [f_r(h,t) - f_r(\hat{h},\hat{t}) + \gamma]_+$$

for which $f_r(h,t) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2$ is the dissimilarity measure of a triple (h,r,t), γ is a positive margin, $[x]_+$ denotes the positive part of x (i.e. $\max(x,0)$), and $(\hat{h}, r, \hat{t})$ is a Bernoulli negative-sampled triple [Wang *et al.*, 2014] by substituting either h or t in (h,r,t).

On top of that, an alignment model jointly captures cross-lingual inferences across language-specific embedding spaces. As previously mentioned, various alignment techniques have been adopted by previous models, among which we choose the linear-transformation-based technique, as we find it offers the best performance in modeling cross-lingual inferences. Hence, the objective function is given as below.

$$S_A = \sum_{(e,e')\in I(L_i,L_j)} \|\mathbf{M}_{ij}\mathbf{e} - \mathbf{e}'\|_2$$

$\mathbf{M}_{ij}$ thereof is a $k_1 \times k_1$ matrix that serves as a linear transformation on entity vectors from L_i to L_j.

The objective of the KGEM component is to minimize $S_{KG} = S_K + \alpha S_A$, for which α is a positive hyperparameter. Conceptually, KGEM is equivalent to a modification of MTransE-LT, where the alignment model is refined from the triple level to the entity level. Meanwhile, the linear transformation $\mathbf{M}_{ij}$ also applies to relation embeddings. Consider

two cross-lingual counterparts of triples (h,r,t), (h',r',t'), since S_K requires $\mathbf{h} + \mathbf{r} \approx \mathbf{t}$ and $\mathbf{h}' + \mathbf{r}' \approx \mathbf{t}'$, it is easy to get $\mathbf{M}_{ij}\mathbf{r} \approx \mathbf{r}'$ from $\mathbf{M}_{ij}(\mathbf{t} - \mathbf{h}) \approx \mathbf{t}' - \mathbf{h}'$. It is noteworthy that, other techniques such as DistMult and HolE may be employed for the knowledge model as well, but we leave them to future work for two reasons: (1) to facilitate the direct comparison with previous works; (2) these techniques disable the cross-lingual inferences on relations.

Like many KG embedding models [Bordes *et al.*, 2013; 2014; Yang *et al.*, 2015; Chen *et al.*, 2017a], we constrain the l_2-norms of entity embeddings to be 1 during the learning process. We notice that some other works [Wang *et al.*, 2014; Lin *et al.*, 2015] constrain such norms to be less than 1 instead, which we find however, do not prevent the optimization from a trivial solution where all vectors collapse towards zero, hence largely impairing the quality of embeddings.

3.2 Multilingual Entity Description Embeddings

The DEM learns in two stages. An attentive gated recurrent unit encoder (AGRU) is used to encode the multilingual entity descriptions. On top of that, DEM is trained to collocate the description embeddings of cross-lingual counterparts.

Gated Recurrent Unit. The gated recurrent unit (GRU) has been popular in sentence (sequence) encoders in recent works [Jozefowicz *et al.*, 2015], which consists of two types of gates to track the state of sequences without using separated memory cells, i.e. the reset gate $\mathbf{g}_t$ and the update gate $\mathbf{z}_t$. Given the vector representation $\mathbf{x}_t$ of an incoming item x_t from the sequence, GRU updates the current state $\mathbf{s}_t$ as a linear interpolation between the previous state $\mathbf{s}_{t-1}$ and the candidate state $\tilde{\mathbf{s}}_t$ of the new item x_t, which is calculated as below.

$$\mathbf{s}_t = \mathbf{z}_t \odot \tilde{\mathbf{s}}_t + (1 - \mathbf{z}_t) \odot \mathbf{s}_{t-1}$$

The update gate $\mathbf{z}_t$ that balances between the information of the previous sequence and the new item is updated as below,

$$\mathbf{z}_t = \sigma\left(\mathbf{M}_z\mathbf{x}_t + \mathbf{N}_z\mathbf{s}_{t-1} + \mathbf{b}_z\right)$$

where $\mathbf{M}_z$ and $\mathbf{N}_z$ are two weight matrices, $\mathbf{b}_z$ is a bias vector, and σ is the sigmoid function. The candidate state $\tilde{\mathbf{s}}_t$ is calculated similarly to those in a traditional recurrent unit as below, where $\mathbf{M}_s$ and $\mathbf{N}_s$ are two weight matrices, and $\mathbf{b}_s$ is a bias vector,

$$\tilde{\mathbf{s}}_t = \tanh\left(\mathbf{M}_s\mathbf{x}_t + \mathbf{g}_t \odot (\mathbf{N}_s\mathbf{s}_{t-1}) + \mathbf{b}_s\right)$$

The reset gate $\mathbf{g}_t$ thereof, controls how much the information from the past sequence contribute to the candidate state, and is updated as below,

$$\mathbf{g}_t = \sigma\left(\mathbf{M}_g\mathbf{x}_t + \mathbf{N}_g\mathbf{s}_{t-1} + \mathbf{b}_g\right)$$

The above defines a GRU layer which outputs a sequence of hidden state vectors given the input sequence X.

Self-attention. The motivation of importing self-attention is to extract the words that contain shared information across the multilingual descriptions of the same entity, where content details can be inconsistent. Consider the two descriptions of Fig. 1. We expect the encoder to highlight the sentence parts with the important shared information such as *scientist* and *scientifique spécialisé* as well as *the field of astronomy* and

l'astronomie, but rather than inconsistent details such as *specific question or field outside of the scope of Earth* in English. We hence incorporate the self-attention defined in [Kim *et al.*, 2017] to the above GRU layer as below.

$$\mathbf{u}_t = \tanh\left(\mathbf{M}_a \mathbf{s}_t + \mathbf{b}_a\right)$$

$$a_t = \frac{\exp\left(\mathbf{u}_t^\top \mathbf{x}_t\right)}{\sum_{x_i \in X} \exp\left(\mathbf{u}_i^\top \mathbf{x}_i\right)}$$

$$\mathbf{v}_t = |X| a_t \mathbf{u}_t$$

$\mathbf{u}_t$ thereof is a hidden representation of $\mathbf{s}_t$ from the GRU layer. A normalized attention weight a_t is calculated through a softmax function, which measures the importance of item x_t in the GRU encoding of sequence X, and is applied to $\mathbf{u}_t$ to obtain the self-attention output $\mathbf{v}_t$. Note that a coefficient $|X|$ (the length of the input sequence) is applied so as to keep $\mathbf{v}_t$ from losing the original scale.

Multilingual Word Embeddings. To better reflect the semantic information of multilingual entity descriptions from the word level, we use multilingual word embeddings that are capable of collocating similar words in different languages. In detail, we pre-train the cross-lingual Bilbowa [Gouws *et al.*, 2015] word embeddings on the cross-lingual parallel corpora Europarl v7 [Koehn, 2005] and monolingual corpora of Wikipedia dump. After the pre-training, we fix the word embeddings to convert each entity description d_e to a sequence of vectors to be fed into the description encoder.

Learning Objective. We utilize an encoder of two stacked attentive GRU layers to model the descriptions of both languages, which takes the description sequence d_e and produces the embedding from the second-layer outputs. In detail, we apply an affine layer to map the averaged second-layer outputs to a common embedding space for descriptions: $\mathbf{d}_e = \tanh\left(\mathbf{M}_d\left(\frac{1}{|d_e|}\sum_{i=1}^{|d_e|}\mathbf{v}_i^{(2)}\right) + \mathbf{b}_d\right)$. We use the same dimensionality (denoted as k_2) for the output vectors of the second GRU layer $\mathbf{v}_i^{(2)}$ and the description embeddings $\mathbf{d}_e$. Like KG embeddings, we regularize each $\mathbf{d}_e$ as $\|\mathbf{d}_e\|_2 = 1$.

The learning objective of DEM is to maximize the log likelihood of each entity given its cross-lingual counterpart in terms of their description embeddings, which is realized by minimizing the following objective function,

$$S_D = \sum_{(e,e') \in I(L_i, L_j)} -LL_1 - LL_2$$

$$= \sum_{(e,e') \in I(L_i, L_j)} -\log\left(P(e|e')\right) - \log\left(P(e'|e)\right)$$

Similar to [Mikolov *et al.*, 2013b], we adopt negative sampling to obtain the following computationally efficient terms of approximation for LL_1 and LL_2, where $|B_d|$ is the batched sampling size, and U is the distribution of entities.

$$LL_1 = \log\sigma\left(\mathbf{d}_e^\top \mathbf{d}_{e'}\right) + \sum_{k=1}^{|B_d|} \mathbb{E}_{e_k \sim \mathrm{U}\left(e_k \in E_{L_i}\right)}\left[\log\sigma\left(-\mathbf{d}_{e_k}^\top \mathbf{d}_{e'}\right)\right]$$

$$LL_2 = \log\sigma\left(\mathbf{d}_e^\top \mathbf{d}_{e'}\right) + \sum_{k=1}^{|B_d|} \mathbb{E}_{e_k \sim \mathrm{U}\left(e_k \in E_{L_j}\right)}\left[\log\sigma\left(-\mathbf{d}_e^\top \mathbf{d}_{e_k}\right)\right]$$

Algorithm 1: Iterative co-training of KDCoE.

Input: Graphs G_{L_i}, G_{L_j}, descriptions D_{L_i}, D_{L_j}, ILL training set I_{tr}, ILL validation set I_{val}, candidate entities without ILLs $\tilde{E}_{L_i} \in E_{L_i}$, $\tilde{E}_{L_j} \in E_{L_j}$, precision threshold τ on I_{val} for selecting proposed ILLs.

Output: parameters θ for KGEM and DEM

1 **while** *Either* KGEM *or* DEM *does not propose more ILLs* **do**
2   Reinitialize KGEM and DEM;
3   Train KGEM on I_{tr}, G_{L_i}, G_{L_j} until S_{KG} no longer improves on graphs and I_{val};
4   Select max l_2 threshold δ_1, for which the precision of the predictions $(e, \hat{e}')$ by KGEM on I_{val} s.t. $\left\|\mathbf{M}_{ij}\mathbf{e} - \hat{\mathbf{e}}'\right\|_2 < \delta_1$ is higher than τ;
5   **for** $e \in \tilde{E}_{L_i}$ **do**
6        $\hat{\mathbf{e}}' \leftarrow \text{NearestNeighbor}(\mathbf{M}_{ij}\mathbf{e}, L_j)$; /* NN in L_j. */
7   **if** $\left\|\mathbf{M}_{ij}\mathbf{e} - \hat{\mathbf{e}}'\right\|_2 < \delta_1$ **then**
8           $I_{tr} \leftarrow I_{tr} \cup \{(e, \hat{e}')\}$; /* Propose an ILL. */
9           $\tilde{E}_{L_i} \leftarrow \tilde{E}_{L_i} - \{e\}$; $\tilde{E}_{L_j} \leftarrow \tilde{E}_{L_j} - \{\hat{e}'\}$;
10   Train DEM on I_{tr}, D_{L_i}, D_{L_j} until S_D no longer improves on I_{val};
11   Select max l_2 threshold δ_2, for which the precision of the predictions $(e, \hat{e}')$ by DEM on I_{val} s.t. $\left\|\mathbf{d}_e - \mathbf{d}_{\hat{e}'}\right\|_2 < \delta_2$ is higher than τ;
12   **for** $e \in \tilde{E}_{L_i}$ **do**
13        $\mathbf{d}_{\hat{e}'} \leftarrow \text{NearestNeighbor}(\mathbf{d}_e, L_j)$; /* NN in L_j. */
14   **if** $\left\|\mathbf{d}_e - \mathbf{d}_{\hat{e}'}\right\|_2 < \delta_2$ **then**
15           $I_{tr} \leftarrow I_{tr} \cup \{(e, \hat{e}')\}$; /* Propose an ILL. */
16           $\tilde{E}_{L_i} \leftarrow \tilde{E}_{L_i} - \{e\}$; $\tilde{E}_{L_j} \leftarrow \tilde{E}_{L_j} - \{\hat{e}'\}$;

Through optimization of S_D, the encoder is trained towards the goal of maximizing the dot product of each description embedding $\mathbf{d}_e$ and that of its cross-lingual counterpart $\mathbf{d}_{e'}$, and decreasing the dot product of unrelated description embeddings. Since description embeddings are regularized to unit vectors, this process is equivalent to minimizing the l_2-distance between each pair of cross-lingual counterparts (i.e. collocating). To facilitate the sampling-based approximation, we use the stratified negative sharing technique [Chen *et al.*, 2017c]. That is to say, we sample batches of ILLs into B_d. Then based on the 1-to-1 mapping of ILLs, we select negative samples for each e as all entities e_k in the other language from B_d, except for the one that forms the ILL with e.

Note that we have also explored with other forms of description encoders. Single-layer (an affine layer applied to averaged word embeddings) and CNN used in [Xie *et al.*, 2016] to represent monolingual entity descriptions fail to accurately match cross-lingual counterparts by losing the sequential and attentive information. Attentive LSTM encoders perform comparably to AGRU, but are more complex and require more computational resources for training. Adopting bidirectional encoders hinders the performance of our tasks.

3.3 Iterative Co-training

The co-training of the two model components is conducted iteratively on the KG, where a small amount of ILLs is provided for training. At each iteration, the component models alternately take turns of the train-and-propose process. In each turn, the model is first initialized using orthogonal initialization, and optimized using SGD with early-stopping based on a small validation set of ILLs. After training, that model predicts new ILLs for candidate entities that are not involved in any previous ILL. Such a pre-

diction is based on a distance-based strategy, where a new ILL sourced from L_i is suggested by searching the nearest neighbor (NN) within the candidate space of L_j from the transformed entity vector, or from the original description vector. As lower l_2-distances imply more precise inferences of embeddings [Chen *et al.*, 2017a; Zhu *et al.*, 2017; Mikolov *et al.*, 2013b], only the most confident predictions, for which the l_2-distance between the source and the NN falls within a certain threshold, are populated into the training set. The l_2-distance threshold is selected to ensure the prediction precision on the validation set to be above τ, so as to ensure a high estimated precision of proposed new ILLs. Both components repeatedly conduct the above train-and-propose processes, therefore gradually enhance the supervision of cross-lingual learning for each other, until either of the two model components no longer proposes new ILLs. The detailed co-training procedure of KDCoE is given in Algorithm 1.

4 Experiments

In this section, we evaluate KDCoE on two knowledge alignment tasks: cross-lingual entity alignment and zero-shot alignment. We also conduct an experiment on cross-lingual KG completion, which aims at enhancing the traditional monolingual KG completion with cross-lingual knowledge.

Dataset. Experiments are conducted on the trilingual dataset WK3l60k, which is extracted from the subset of DBpedia that is highly covered by ILLs in the purpose of providing enough ground truth to evaluate the semi-supervised cross-lingual learning. Statistics of the dataset is given in Table 1. Each language-specific version of the KG consists of 54k to 65k entities, and varies in density, which indicates the dataset to be challenging in terms of cross-lingual inconsistency and providing much larger candidate spaces than other datasets for KG embeddings that typically searches around 15k-40k entities [Yang *et al.*, 2015; Sun *et al.*, 2017]. Literal descriptions covers 82%-96% of entities in each language. We extract ILLs between English-French and English-German to train and evaluate cross-lingual entity alignment, for which we use about 20% for training, 70% for testing, and the rest for validation. The proportion used for training is in accord with the estimated global completeness of ILLs in the KB [Lehmann *et al.*, 2015]. Meanwhile, another small set of entities with ILLs and descriptions are extracted, but are excluded from the KG structure for evaluating zero-shot alignment.

4.1 Cross-lingual Entity Alignment

The objective of this task is to match the same entities from different languages in KB. The baselines we compare against include three MTransE variants that adopt different alignment techniques to model ILLs, and ITransE which employs parameter sharing for self-training. We also adapt LM, CCA, and OT (as introduced in Section 2) to their KG equivalences.

Evaluation Protocol. The MTransE variants, ITransE, and KGEM of KDCoE are trained on the complete KG structures of two languages and the small training set of ILLs. LM and CCA are implemented by inducing the corresponding transformations across separately trained knowledge models. OT

Data	#En	#Fr	#De	ILL Lang	#Train	#Valid	#Test	#Zero-shot
Triples	569,393	258,337	224,647	En-Fr	13,050	2,000	39,155	5,000
Desc.	67,314	45,842	43,559	En-De	12,505	2,000	41,018	5,632

Table 1: Statistics of the Wk3l60k dataset.

Language	En-Fr			En-De		
Metric	$Hit@1$	$Hit@10$	MRR	$Hit@1$	$Hit@10$	MRR
LM	1.02	2.21	0.014	1.37	2.14	0.015
CCA	1.80	3.54	0.021	2.19	3.42	0.025
OT	20.15	25.37	0.212	11.04	19.74	0.122
ITransE	10.14	11.59	0.106	6.55	11.44	0.076
MTransE-AC	4.49	8.67	0.051	5.56	8.50	0.060
MTransE-TV	5.12	7.55	0.055	3.62	8.12	0.053
MTransE-LT	27.40	33.98	0.309	17.90	31.59	0.225
KDCoE ($i2$)	37.70	45.01	0.405	29.80	41.66	0.322
KDCoE ($i3$)	43.77	53.07	0.463	30.99	43.02	0.334
KDCoE ($i4$)	46.17	54.85	0.487	32.20	44.58	0.346
KDCoE (term)	**48.32**	**56.95**	**0.496**	**33.52**	**45.47**	**0.349**

Table 2: Results of cross-lingual entity alignment.

is implemented by enforcing MTransE-LT with an orthogonality constraint. DEM of KDCoE is trained on the entity descriptions that are covered by the current I_{tr} during each iteration of co-training. For each ILL (e, e'), the prediction is performed by a kNN search from the cross-lingual conversion point of $\mathbf{e}$, and record the rank of $\mathbf{e'}$ within related entities in the target language. Following the convention [Nickel *et al.*, 2016], we aggregate three metrics on test cases: the accuracy $Hit@1$ (%), the proportion of ranks no larger than 10 $Hit@10$ (%), and mean reciprocal rank MRR. All three metrics are preferred to be higher to indicate better performance.

Model configuration is based on the validation set. We search the learning rate λ_1 for KGEM and other baselines among $\{0.001, 0.005, 0.01\}$, dimensionality k_1 in $\{50, 75, 100\}$, margin γ in $\{0.5, 1, 2\}$, and α in $\{1, 2.5, 5, 7.5\}$. For ITransE, we select the distance threshold δ_i for self-training among $\{0.5, 0.75, 1\}$. For DEM of KDCoE we select the learning rate λ_2 among $\{0.001, 0.005, 0.01\}$, dimensionality k_2 in $\{50, 75, 100\}$. We fix the batch sizes $|B_t|$ for KGEM and other models, and $|B_d|$ for DEM as 1024. The best configuration is $\lambda_1 = 0.005$, $k_1 = 50$, $\gamma = 1$, $\alpha = 2.5$, $\delta_i = 0.75$ for all KG embedding models, and $\lambda_2 = 0.001$, $k_2 = 75$ for DEM. For ILL proposing, we set the precision threshold τ to 0.9. We pre-train Bilbowa based on the setting in [Gouws *et al.*, 2015] to obtain 200-dimensional word embeddings. The multilingual entity descriptions are delimited to the first two sentences, so as to reduce some inconsistent content details. We also remove the stop words in these descriptions, zero-pad short ones and truncate long ones to the average sequence length of 36. Training of models is always terminated via early-stopping, and the co-training process of KDCoE is terminated when either component is not able to propose ILLs for at least 1% of the entity vocabulary.

Results. Results are reported in Table 2, where the results by KDCoE are reported for three co-training iterations since the second iteration where KGEM is first leveraged, and for its final stage (which are respectively marked as KDCoE ($i2 - i4$) and KDCoE (term)). Among all baselines, MTransE-LT notably outperforms others, including other MTransE variants. The orthogonality constraint of OT seems to be too strict so

Language	En-Fr			En-De		
Metric	$Hit@1$	$Hit@10$	MRR	$Hit@1$	$Hit@10$	MRR
Single-layer	0.97	1.80	0.013	0.36	2.10	0.010
CNN	1.19	6.91	0.036	1.28	4.63	0.019
GRU	18.45	27.65	0.204	11.23	24.48	0.165
AGRU-mono	5.08	18.27	0.096	5.03	14.90	0.085
AGRU-multi	26.92	44.69	0.337	19.34	45.69	0.269
KDCoE ($i1$)	27.69	48.69	0.346	19.52	45.84	0.274
KDCoE ($i2$)	28.82	52.58	0.350	20.37	46.35	0.279
KDCoE ($i3$)	30.83	55.91	**0.384**	21.28	48.49	0.283
KDCoE (term)	**30.96**	**56.93**	0.382	**21.97**	**50.02**	**0.285**

Table 3: Results of zero-shot alignment.

Language	Fr				De			
Predict	Tail		Head		Tail		Head	
Metric	$Hit@10$	MRR	$Hit@10$	MRR	$Hit@10$	MRR	$Hit@10$	MRR
TransE	29.21	0.077	18.19	0.046	29.58	0.099	23.57	0.059
KDCoE-mono	31.05	0.092	16.88	0.053	29.13	0.124	27.63	0.106
KDCoE-cross	**37.21**	**0.139**	**22.23**	**0.093**	**34.17**	**0.134**	**31.05**	**0.143**

Table 4: Results of KG completion.

that it impairs the performance. ITransE works well on aligning coherent monolingual KGs [Zhu *et al.*, 2017], but does not adapt well to the inconsistent multilingual KGs. Without jointly adapting the monolingual vector spaces with the alignment, off-line approaches LM and CCA are left behind. On both language settings, KDCoE is able to gradually improve MTransE-LT in every iteration of co-training. The most significant improvements happen in the first iterations, where a majority of candidate ILLs are to be proposed. The final stages of KDCoE (6th and 5th iterations of the two settings) outperform the best baseline by almost doubling $Hit@1$ as well as offering significantly higher $Hit@10$ and MRR. Hence, the co-training approach of KDCoE on enhancing semi-supervised entity alignment is very promising.

4.2 Zero-shot Alignment

This task focuses on aligning entities that do not exist in the structure of KG. While existing KG embedding models require candidates to occur for at least once in the KG structures, KDCoE is capable of dealing with zero-shot scenarios based on the representations of descriptions. For this task, we evaluate KDCoE by aligning the *zero-shot set* of WK3l60k, which are excluded from the KG structures for training. Meanwhile, we also compare the vanilla AGRU without co-training (AGRU-multi) against other encoding techniques, so as to show the effectiveness of our DEM. These baselines include the Single-layer encoder that applies an affine layer to the averaged word embeddings of a description and the two-layer CNN with max-pooling in [Xie *et al.*, 2016] that have been used to encode monolingual descriptions, as well as a two-layer GRU encoder without attention. We also substitute Bilbowa with monolingual Skipgram [Mikolov *et al.*, 2013b] in AGRU (AGRU-mono) so as to verify the effectiveness of incorporating multilingual word embeddings.

Evaluation Protocol. We carry forward the corresponding configurations from the last experiment to show the performance under controlled variables. Specifically for CNN, we follow [Xie *et al.*, 2016] to use 4-max-pooling and kernel-size of 2. Skipgram is trained separately on Wikipedia dumps

of two languages towards 200-dimensional word vectors for AGRU-mono. All the baselines are trained on the ILL training set and corresponding descriptions. The results of KDCoE are reported for the first three iterations and the final stage.

Results. Results in Table 3 show that the vanilla DEM of AGRU outperforms the other encoders. This also indicates that AGRU is more competent for proposing ILLs in co-training based on unseen descriptions than others. As expected, co-training effectively leverages the zero-shot alignment with an increment of $Hit@1$ by 4.04% and 2.63%, as well as $Hit@10$ by 11.97% and 4.33% respectively on the two language settings. The results by GRU and AGRU-mono show that self-attention and multilingual word embeddings are vital to capture the cross-lingual semantic relatedness of descriptions from the word level. Failing to capture the sequence information, Single-layer and CNN are left behind.

4.3 Cross-lingual KG Completion

Lastly, we compare the KGEM of KDCoE against its monolingual counterpart TransE for KG completion, based on the sparser French and German versions of WK3l60k. We explore with two prediction methods for KDCoE. *Monolingual prediction* (KDCoE-mono) aims to query the missing h or t of a triple (h, r, t) in the same way of TransE by searching among the entities of the same language to minimize the dissimilarity function $f_r(h, t)$ (Section 3.2). *Cross-lingual prediction* (KDCoE-cross) provides a new method of triple completion, by converting the monolingual prediction process to the embedding space of another language, then convert the results back to the source language. The idea of cross-lingual prediction is to leverage the traditional monolingual KG completion using a well-populated KG structure of an intermediary language given limited cross-lingual alignment.

Evaluation Protocol. We hold-out 10k French and German triples as test data. KDCoE is co-trained on the rest of the training data till termination. Cross-lingual predictions are processed in the space of English. TransE follows the configuration of KGEM in the previous experiments, and is trained on the KG structure of each language excluding the test data.

Results. The results for $Hit@10$ and MRR are reported in Table 4. KDCoE-mono performs at least comparably to TransE, which indicates that KDCoE preserves well the characterization of monolingual KG structures. Meanwhile, results of cross-lingual prediction prove feasibility of this new method by offering noticeably better outcomes than monolingual prediction. Although this experiment is relatively simple, and may subject to the adequacy of knowledge in the intermediary language, this method opens up a new direction of future work for this task. Moreover, suppose more languages of KGs are provided, we are interested in exploring an ensemble approach [Chen and Guestrin, 2016] that interpolates multiple KDCoEs on different bridges of languages to co-populate one sparse language-specific version of KG.

5 Conclusion and Future Work

In this paper, we propose a semi-supervised learning approach to co-train multilingual KG embeddings and the embeddings of entity descriptions for cross-lingual knowledge

alignment. Our approach `KDCoE` effectively leverages KG embeddings for learning cross-lingual inferences on large, weakly-aligned KGs, which significantly outperforms previous models on the entity alignment task. The zero-shot alignment task also shows the effectiveness of `KDCoE` for improving the cross-lingual matching of entity descriptions through co-training. Meanwhile, we observe that `KDCoE` is able to enhance the traditional methods of KG completion by leveraging the information from another language. For future work, besides the boosting approach mentioned in Section 4.3 for cross-lingual KG completion, we seek to explore the effect of other forms of knowledge models in KGEM for encoding each language-specific KG structure.

References

[Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, et al. Translating embeddings for modeling multi-relational data. In *NIPS*, 2013.

[Bordes *et al.*, 2014] Antoine Bordes, Jason Weston, et al. Open question answering with weakly supervised embedding models. In *ECML-PKDD*, 2014.

[Chen and Guestrin, 2016] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *KDD*, 2016.

[Chen *et al.*, 2017a] Muhao Chen, Yingtao Tian, et al. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *IJCAI*, 2017.

[Chen *et al.*, 2017b] Muhao Chen, Tao Zhou, et al. Multigraph affinity embeddings for multilingual knowledge graphs. In *AKBC*, 2017.

[Chen *et al.*, 2017c] Ting Chen, Yizhou Sun, et al. On sampling strategies for neural network-based collaborative filtering. In *KDD*, 2017.

[Chen *et al.*, 2018] Muhao Chen, Yingtao Tian, et al. On2vec: Embedding-based relation prediction for ontology population. In *SDM*, 2018.

[Dettmers *et al.*, 2018] Tim Dettmers, Pasquale Minervini, et al. Convolutional 2d knowledge graph embeddings. In *AAAI*, 2018.

[Fang *et al.*, 2017] Yuan Fang, Kingsley Kuan, et al. Object detection meets knowledge graphs. In *IJCAI*, 2017.

[Faruqui and Dyer, 2014] Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. *EACL*, 2014.

[Gouws *et al.*, 2015] Stephan Gouws, Yoshua Bengio, et al. Bilbowa: Fast bilingual distributed representations without word alignments. In *ICML*, 2015.

[He *et al.*, 2017] He He, Anusha Balakrishnan, et al. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *ACL*, 2017.

[Ji *et al.*, 2015] Guoliang Ji, Shizhu He, et al. Knowledge graph embedding via dynamic matrix. In *ACL*, 2015.

[Jia *et al.*, 2016] Yantao Jia, Yuanzhuo Wang, et al. Locally adaptive translation for knowledge graph embedding. In *AAAI*, 2016.

[Jozefowicz *et al.*, 2015] Rafal Jozefowicz, Wojciech Zaremba, et al. An empirical exploration of recurrent network architectures. In *ICML*, 2015.

[Kim *et al.*, 2017] Yoon Kim, Carl Denton, et al. Structured attention networks. In *ICLR*, 2017.

[Koehn, 2005] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, 2005.

[Lehmann *et al.*, 2015] Jens Lehmann, Robert Isele, et al. Dbpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 2015.

[Lin *et al.*, 2015] Yankai Lin, Zhiyuan Liu, et al. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, 2015.

[Mahdisoltani *et al.*, 2015] Farzaneh Mahdisoltani, Joanna Biega, et al. Yago3: A knowledge base from multilingual Wikipedias. In *CIDR*, 2015.

[Mikolov *et al.*, 2013a] Tomas Mikolov, Quoc V Le, et al. Exploiting similarities among languages for machine translation. *CoRR*, 2013.

[Mikolov *et al.*, 2013b] Tomas Mikolov, Ilya Sutskever, et al. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[Nickel *et al.*, 2016] Maximilian Nickel, Lorenzo Rosasco, et al. Holographic embeddings of knowledge graphs. In *AAAI*, 2016.

[Speer *et al.*, 2017] Robert Speer, Joshua Chin, et al. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 2017.

[Sun *et al.*, 2017] Zequn Sun, Wei Hu, et al. Cross-lingual entity alignment via joint attribute-preserving embedding. In *ISWC*, 2017.

[Thi *et al.*, 2016] Do Thi, Ngoc Quynh, et al. Facing the most difficult case of semantic role labeling: A collaboration of word embeddings and co-training. In *ACL*, 2016.

[Wan, 2009] Xiaojun Wan. Co-training for cross-lingual sentiment classification. In *ACL-IJCNLP*, 2009.

[Wang *et al.*, 2014] Zhen Wang, Jianwen Zhang, et al. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, 2014.

[Xie *et al.*, 2016] Ruobing Xie, Zhiyuan Liu, et al. Representation learning of knowledge graphs with entity descriptions. In *AAAI*, 2016.

[Xing *et al.*, 2015] Chao Xing, Dong Wang, et al. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL*, 2015.

[Yang *et al.*, 2015] Bishan Yang, Wen-tau Yih, et al. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*, 2015.

[Zhang *et al.*, 2014] Mi Zhang, Jie Tang, et al. Addressing cold start in recommender systems: A semi-supervised co-training algorithm. In *SIGIR*, 2014.

[Zhu *et al.*, 2017] Hao Zhu, Ruobing Xie, et al. Iterative entity alignment via knowledge embeddings. In *IJCAI*, 2017.