

Multilingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment

Muhao Chen¹, Yingtao Tian², Mohan Yang¹, Carlo Zaniolo¹

Department of Computer Science, UCLA¹

Department of Computer Science, Stony Brook University²

{muhaochen, yang, zaniolo}@cs.ucla.edu; yittian@cs.stonybrook.edu

Abstract

Many recent works have demonstrated the benefits of knowledge graph embeddings in completing monolingual knowledge graphs. Inasmuch as related knowledge bases are built in several different languages, achieving cross-lingual knowledge alignment will help people in constructing a coherent knowledge base, and assist machines in dealing with different expressions of entity relationships across diverse human languages. Unfortunately, achieving this highly desirable cross-lingual alignment by human labor is very costly and error-prone. Thus, we propose MTransE, a translation-based model for multilingual knowledge graph embeddings, to provide a simple and automated solution. By encoding entities and relations of each language in a separated embedding space, MTransE provides transitions for each embedding vector to its cross-lingual counterparts in other spaces, while preserving the functionalities of monolingual embeddings. We deploy three different techniques to represent cross-lingual transitions, namely axis calibration, translation vectors, and linear transformations, and derive five variants for MTransE using different loss functions. Our models can be trained on partially aligned graphs, where just a small portion of triples are aligned with their cross-lingual counterparts. The experiments on cross-lingual entity matching and triple-wise alignment verification show promising results, with some variants consistently outperforming others on different tasks. We also explore how MTransE preserves the key properties of its monolingual counterpart TransE.

1 Introduction

Multilingual knowledge bases such as Wikipedia [Wikipedia, 2017], WordNet [Bond and Foster, 2013], and ConceptNet [Speer and Havasi, 2013] are becoming essential sources of knowledge for people and AI-related applications. These knowledge bases are modeled as knowledge graphs that store two aspects of knowledge: the *monolingual knowledge* that includes entities and relations recorded in the form of triples,

and the *cross-lingual knowledge* that matches the monolingual knowledge among various human languages.

The coverage issue of monolingual knowledge has been widely addressed, and parsing-based techniques for completing monolingual knowledge bases have been well studied in the past [Culotta and Sorensen, 2004; Zhou *et al.*, 2005; Sun *et al.*, 2011]. More recently, much attention has been paid to embedding-based techniques, which provide simple methods to encode entities in low-dimensional embedding spaces and capture relations as means of translations among entity vectors. Given a triple (h, r, t) where r is the relation between entities h and t , then h and t are represented as two k -dimensional vectors \mathbf{h} and \mathbf{t} , respectively. A function $f_r(\mathbf{h}, \mathbf{t})$ is used to measure the plausibility of (h, r, t) , which also implies the transformation \mathbf{r} that characterizes r . Exemplarily, the translation-based model TransE [Bordes *et al.*, 2013] uses the loss function $f_r(\mathbf{h}, \mathbf{t}) = \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|^1$, where \mathbf{r} is characterized as a translation vector learnt from the latent connectivity patterns in the knowledge graph. This model provides a flexible way of predicting a missing item in a triple, or verifying the validity of a generated triple. Other works like TransH [Wang *et al.*, 2014] and TransR [Lin *et al.*, 2015], introduce different loss functions that represent the relational translation in other forms, and have achieved promising results in completing the knowledge graphs.

While embedding-based techniques can help improve the completeness of monolingual knowledge, the problem of applying these techniques on cross-lingual knowledge remains largely unexplored. Such knowledge, including *inter-lingual links* (ILLs) that match the same entities, and *triple-wise alignment* (TWA) that represents the same relations, is very helpful in synchronizing different language-specific versions of a knowledge base that evolve independently, as needed to further improve applications built on knowledge bases, such as Q&A systems, semantic Web, and Web search. In spite of its importance, this cross-lingual knowledge remains largely intact. In fact, in the most successful knowledge base Wikipedia, we find that ILLs cover less than 15% entity alignment.

Leveraging knowledge graph embeddings to cross-lingual knowledge no doubt provides a generic way to help extract and apply such knowledge. However, it is a non-trivial task

¹Hereafter, $\|\cdot\|$ means l_1 or l_2 norm unless explicitly specified.

to find a tractable technique to capture the cross-lingual transitions². Such transitions are more difficult to capture than relational translations for several reasons: (i) a cross-lingual transition has a far larger domain than any monolingual relational translation; (ii) it applies on both entities and relations, which have incoherent vocabularies among different languages; (iii) the known alignment for training such transitions usually accounts for a small percentage of a knowledge base. Moreover, the characterization of monolingual knowledge graph structures has to be well-preserved to ensure the correct representation of the knowledge to be aligned.

To address the above issues, we propose a multilingual knowledge graph embedding model *MTransE*, that learns the multilingual knowledge graph structure using a combination of two component models, namely *knowledge model* and *alignment model*. The knowledge model encodes entities and relations in a language-specific version of knowledge graph. We explore the method that organizes each language-specific version in a separated embedding space, in which *MTransE* adopts TransE as the knowledge model. On top of that, the alignment model learns cross-lingual transitions for both entities and relations across different embedding spaces, where the following three representations of cross-lingual alignment are considered: distance-based axis calibration, translation vectors, and linear transformations. Thus, we obtain five variants of *MTransE* based on different loss functions, and identify the best variant by comparing them on cross-lingual alignment tasks using two partially aligned trilingual graphs constructed from Wikipedia triples. We also show that *MTransE* performs as well as its monolingual counterpart TransE on monolingual tasks.

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

While, at the best of our knowledge, there is no previous work on learning multilingual knowledge graph embeddings, we will describe next three lines of work which are closely related to this topic.

Knowledge Graph Embeddings. Recently, significant advancement has been made in using the translation-based method to train monolingual knowledge graph embeddings. To characterize a triple (h, r, t) , models of this family follow a common assumption $\mathbf{h}_r + \mathbf{r} \approx \mathbf{t}_r$, where \mathbf{h}_r and \mathbf{t}_r are either the original vectors of h and t , or the transformed vectors under a certain transformation w.r.t. relation r . The forerunner TransE [Bordes *et al.*, 2013] sets \mathbf{h}_r and \mathbf{t}_r as the original \mathbf{h} and \mathbf{t} , and achieves promising results in handling 1-to-1 relations. Later works improve TransE on multi-mapping relations by introducing relation-specific transformations on entities to obtain different \mathbf{h}_r and \mathbf{t}_r , including projections on relation-specific hyperplanes in TransH [Wang *et al.*, 2014], linear transformations to heterogeneous relation spaces in TransR [Lin *et al.*, 2015], dynamic matrices in

TransD [Ji *et al.*, 2015], and other forms [Jia *et al.*, 2016; Nguyen *et al.*, 2016]. All these variants of TransE specialize entity embeddings for different relations, therefore improving knowledge graph completion on multi-mapping relations at the cost of increased model complexity. Meanwhile translation-based models cooperate well with other models. For example, variants of TransE are combined with word embeddings to help relation extraction from text [Weston *et al.*, 2013; Zhong *et al.*, 2015].

In addition to these, there are non-translation-based methods. Some of those including UM [Bordes *et al.*, 2011], SE [Bordes *et al.*, 2012], Bilinear [Jenatton *et al.*, 2012], and HoIE [Nickel *et al.*, 2016], do not explicitly represent relation embeddings. Others including neural-based models SLM [Collobert and Weston, 2008] and NTN [Socher *et al.*, 2013], and random-walk-based model TADW [Yang *et al.*, 2015a], are expressive and adaptable for both structured and text corpora, but are too complex to be incorporated into an architecture supporting multilingual knowledge.

Multilingual Word Embeddings. Several approaches learn multilingual word embeddings on parallel text corpora. Some of those can be extended to multilingual knowledge graphs, such as LM [Mikolov *et al.*, 2013] and CCA [Faruqui and Dyer, 2014] which induce offline transitions among pre-trained monolingual embeddings in forms of linear transformations and canonical correlation analysis respectively. These approaches do not adjust the inconsistent vector spaces via calibration or jointly training with the alignment model, thus fail to perform well on knowledge graphs as the parallelism exists only in small portions. A better approach OT [Xing *et al.*, 2015] jointly learns regularized embeddings and orthogonal transformations, which is however found to be overcomplicated due to the inconsistency of monolingual vector spaces and the large diversity of relations among entities.

Knowledge Bases Alignment. Some projects produce cross-lingual alignment in knowledge bases at the cost of extensive human involvement and designing hand-crafted features dedicated to specific applications. Wikidata [Vrandečić, 2012] and DBpedia [Lehmann *et al.*, 2015] rely on crowdsourcing to create ILLs and relation alignment. YAGO [Mahdisoltani *et al.*, 2015] mines association rules on known matches, which combines many confident scores and requires extensively fine tuning. Many other works require sources that are external to the graphs, from well-established schemata or ontologies [Nguyen *et al.*, 2011; Suchanek *et al.*, 2011; Rinser *et al.*, 2013] to entity descriptions [Yang *et al.*, 2015b], which being unavailable to many knowledge bases such as YAGO, WordNet, and ConceptNet [Speer and Havasi, 2013]. Such approaches also involve complicated model dependencies that are not tractable and reusable. By contrast, embedding-based methods are simple and general, require little human involvement, and generate task-independent features that can contribute to other NLP tasks.

3 Multilingual Knowledge Graph Embeddings

We hereby begin our modeling with the formalization of multilingual knowledge graphs.

²We use the word *transition* here to differentiate from the relational translations among entities in translation-based methods.

3.1 Multilingual Knowledge Graphs

In a knowledge base KB , we use \mathcal{L} to denote the set of languages, and \mathcal{L}^2 to denote the 2-combination of \mathcal{L} (i.e., the set of *unordered* language pairs). For a language $L \in \mathcal{L}$, G_L denotes the language-specific knowledge graph of L , and E_L and R_L respectively denote the corresponding vocabularies of entity expression and relation expression. $T = (h, r, t)$ 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} respectively represent the embedding vectors of head h , relation r , and tail t . For a language pair $(L_1, L_2) \in \mathcal{L}^2$, $\delta(L_1, L_2)$ denotes the alignment set which contains the pairs of triples that have already been aligned between L_1 and L_2 . For example, across the languages English and French, we may have $((\text{State of California, capital city, Sacramento}), (\text{État de Californie, capitale, Sacramento})) \in \delta(\text{English, French})$. The alignment set commonly exists in a small portion in a multilingual knowledge base [Vrandečić, 2012; Mahdisoltani *et al.*, 2015; Lehmann *et al.*, 2015], and is one part of knowledge we want to extend.

Our model consists of two components that learn on the two facets of KB : the knowledge model that encodes the entities and relations from each language-specific graph structure, and the alignment model that learns the cross-lingual transitions from the existing alignment. We define a model for each language pair from \mathcal{L}^2 that has a non-empty alignment set. Thus, for a KB with more than two languages, a set of models composes the solution. In the following, we use a language pair $(L_i, L_j) \in \mathcal{L}^2$ as an example to describe how we define each component of a model.

3.2 Knowledge Model

For each language $L \in \mathcal{L}$, a dedicated k -dimensional embedding space \mathbb{R}_L^k is assigned for vectors of E_L and R_L , where \mathbb{R} is the field of real numbers. We adopt the basic translation-based method of TransE for each involved language, which benefits the cross-lingual tasks by representing embeddings uniformly in different contexts of relations. Therefore its loss function is given below:

$$S_K = \sum_{L \in \{L_i, L_j\}} \sum_{(h, r, t) \in G_L} \|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$$

It measures the plausibility of all given triples. By minimizing the loss function, the knowledge model preserves monolingual relations among entities, while also acts as a regularizer for the alignment model. Meanwhile, the knowledge model partitions the knowledge base into disjoint subsets that can be trained in parallel.

3.3 Alignment Model

The objective of the alignment model is to construct the transitions between the vector spaces of L_i and L_j . Its loss function is given as below:

$$S_A = \sum_{(T, T') \in \delta(L_i, L_j)} S_a(T, T')$$

for which the alignment score $S_a(T, T')$ iterates through all pairs of aligned triples. Three different techniques to score

the alignment are considered: distance-based axis calibration, translation vectors, and linear transformations. Each of them is based on a different assumption, and constitutes different forms of S_a alongside.

Distance-based Axis Calibration. This type of alignment models penalize the alignment based on the distances of cross-lingual counterparts. Either of the following two scorings can be adopted to the model.

$$S_{a_1} = \|\mathbf{h} - \mathbf{h}'\| + \|\mathbf{t} - \mathbf{t}'\|$$

S_{a_1} regulates that correctly aligned multilingual expressions of the same entity tend to have close embedding vectors. Thus by minimizing the loss function that involves S_{a_1} on known pairs of aligned triples, the alignment model adjusts axes of embedding spaces towards the goal of coinciding the vectors of the same entity in different languages.

$$S_{a_2} = \|\mathbf{h} - \mathbf{h}'\| + \|\mathbf{r} - \mathbf{r}'\| + \|\mathbf{t} - \mathbf{t}'\|$$

S_{a_2} overlays the penalty of relation alignment to S_{a_1} to explicitly converge coordinates of the same relation.

The alignment models based on axis calibration assume analogous spatial emergence of items in each language. Therefore, it realizes the cross-lingual transition by carrying forward the vector of a given entity or relation from the space of the original language to that of the other language.

Translation Vectors. This model encodes cross-lingual transitions into vectors. It consolidates alignment into graph structures and characterizes cross-lingual transitions as regular relational translations. Hence S_{a_3} as below is derived.

$$S_{a_3} = \|\mathbf{h} + \mathbf{v}_{ij}^e - \mathbf{h}'\| + \|\mathbf{r} + \mathbf{v}_{ij}^r - \mathbf{r}'\| + \|\mathbf{t} + \mathbf{v}_{ij}^e - \mathbf{t}'\|$$

\mathbf{v}_{ij}^e and \mathbf{v}_{ij}^r thereof are respectively deployed as the entity-dedicated and relation-dedicated translation vectors between L_i and L_j , such that we have $\mathbf{e} + \mathbf{v}_{ij}^e \approx \mathbf{e}'$ for embedding vectors \mathbf{e} , \mathbf{e}' of the same entity e expressed in both languages, and $\mathbf{r} + \mathbf{v}_{ij}^r \approx \mathbf{r}'$ for those of the same relation. We deploy two translation vectors instead of one, because there are far more distinct entities than relations, and using one vector easily leads to imbalanced signals from relations.

Such a model obtains a cross-lingual transition of an embedding vector by adding the corresponding translation vector. Moreover, it is easy to see that $\mathbf{v}_{ij}^e = -\mathbf{v}_{ji}^e$ and $\mathbf{v}_{ij}^r = -\mathbf{v}_{ji}^r$ hold. Therefore, as we obtain the translation vectors from L_i to L_j , we can always use the same vectors to translate in the opposite direction.

Linear Transformations. The last category of alignment models deduce linear transformations between embedding spaces. S_{a_4} as below learns a $k \times k$ square matrix \mathbf{M}_{ij}^e as a linear transformation on entity vectors from L_i to L_j , given k as the dimensionality of the embedding spaces.

$$S_{a_4} = \|\mathbf{M}_{ij}^e \mathbf{h} - \mathbf{h}'\| + \|\mathbf{M}_{ij}^e \mathbf{t} - \mathbf{t}'\|$$

S_{a_5} additionally brings in a second linear transformation \mathbf{M}_{ij}^r for relation vectors, which is of the same shape as \mathbf{M}_{ij}^e . The use of a different matrix is again due to different redundancy of entities and relations.

$$S_{a_5} = \|\mathbf{M}_{ij}^e \mathbf{h} - \mathbf{h}'\| + \|\mathbf{M}_{ij}^r \mathbf{r} - \mathbf{r}'\| + \|\mathbf{M}_{ij}^e \mathbf{t} - \mathbf{t}'\|$$

Table 1: Summary of model variants.

Var	Model Complexity	Cross-lingual Transition	Search Complexity
Var ₁	$O(n_e kl + n_r kl)$	$\tau_{ij}(\mathbf{e}) = \mathbf{e}$ $\tau_{ij}(\mathbf{r}) = \mathbf{r}$	$O(n_e k)$ $O(n_r k)$
Var ₂	$O(n_e kl + n_r kl)$	$\tau_{ij}(\mathbf{e}) = \mathbf{e}$ $\tau_{ij}(\mathbf{r}) = \mathbf{r}$	$O(n_e k)$ $O(n_r k)$
Var ₃	$O(n_e kl + n_r kl + kl^2)$	$\tau_{ij}(\mathbf{e}) = \mathbf{e} + \mathbf{v}_{ij}^e$ $\tau_{ij}(\mathbf{r}) = \mathbf{r} + \mathbf{v}_{ij}^r$	$O(n_e k)$ $O(n_r k)$
Var ₄	$O(n_e kl + n_r kl + 0.5k^2 l^2)$	$\tau_{ij}(\mathbf{e}) = \mathbf{M}_{ij}^e \mathbf{e}$ $\tau_{ij}(\mathbf{r}) = \mathbf{M}_{ij}^r \mathbf{r}$	$O(n_e k^2 + n_e k)$ $O(n_r k^2 + n_r k)$
Var ₅	$O(n_e kl + n_r kl + k^2 l^2)$	$\tau_{ij}(\mathbf{e}) = \mathbf{M}_{ij}^e \mathbf{e}$ $\tau_{ij}(\mathbf{r}) = \mathbf{M}_{ij}^r \mathbf{r}$	$O(n_e k^2 + n_e k)$ $O(n_r k^2 + n_r k)$

Notation: \mathbf{e} and \mathbf{r} are respectively the vectors of an entity e and a relation r , k is the dimension of the embedding spaces, l is the cardinality of \mathcal{L} , n_e and n_r are respectively the number of entities and the number of relations, where $n_e \gg n_r$.

Unlike axis calibration, linear-transformation-based alignment model treats cross-lingual transitions as the topological transformation of embedding spaces without assuming the similarity of spatial emergence.

The cross-lingual transition of a vector is obtained by applying the corresponding linear transformation. It is noteworthy that, regularization of embedding vectors in the training process (which will be introduced soon after) ensures the invertibility of the linear transformations such that $\mathbf{M}_{ij}^{e,-1} = \mathbf{M}_{ji}^e$ and $\mathbf{M}_{ij}^{r,-1} = \mathbf{M}_{ji}^r$. Thus the transition in the revert direction is always enabled even though the model only learns the transformations of one direction.

3.4 Variants of MTransE

Combining the above two component models, MTransE minimizes the following loss function $J = S_K + \alpha S_A$, where α is a hyperparameter that weights S_K and S_A .

As we have given out five variants of the alignment model, each of which correspondingly defines its specific way of computing cross-lingual transitions of embedding vectors. We denote Var _{k} as the variant of MTransE that adopts the k -th alignment model which employs S_{a_k} . In practice, the searching of a cross-lingual counterpart for a source is always done by querying the nearest neighbor from the result point of the cross-lingual transition. We denote function τ_{ij} that maps a cross-lingual transition of a vector from L_i to L_j , or simply τ in a bilingual context. As stated, the solution in a multi-lingual scenario consists of a set of models of the same variant defined on every language pair in \mathcal{L}^2 . Table 1 summarizes the model complexity, the definition of cross-lingual transitions, and the complexity of searching a cross-lingual counterpart for each variant.

3.5 Training

We optimize the loss function using on-line stochastic gradient descent [Wilson and Martinez, 2003]. At each step, we update the parameter θ by setting $\theta \leftarrow \theta - \lambda \nabla_{\theta} J$, where λ is the learning rate. Instead of directly updating J , our implementation optimizes S_K and αS_A alternately. In detail, at each epoch we optimize $\theta \leftarrow \theta - \lambda \nabla_{\theta} S_K$ and $\theta \leftarrow \theta - \lambda \nabla_{\theta} \alpha S_A$ in separated groups of steps.

We enforce the constraint that the l_2 norm of any entity embedding vector is 1, thus regularize embedding vectors to a unit spherical surface. This constraint is employed in the literature [Bordes *et al.*, 2013; 2014; Jenatton *et al.*, 2012] and has two important effects: (i) it helps avoid the case

Table 2: Statistics of the WK31 data sets.

Data set	#En triples	#Fr triples	#De triples	#Aligned triples
WK31-15k	203,502	170,605	145,616	En-Fr:16,470 En-De:37,170
WK31-120k	1,376,011	767,750	391,108	En-Fr:124,433 En-De:69,413

Table 3: Number of entity inter-lingual links (ILLs).

Data Set	En-Fr	Fr-En	En-De	De-En
WK31-15k	3,733	3,815	1,840	1,610
WK31-120k	42,413	41,513	7,567	5,921

where the training process trivially minimizes the loss function by shrinking the norm of embedding vectors, and (ii) it implies the invertibility of the linear transformations [Xing *et al.*, 2015] for Var₄ and Var₅.

We initialize vectors by drawing from a uniform distribution on the unit spherical surface, and initialize matrices using random orthogonal initialization [Saxe *et al.*, 2014]. Negative sampling is not employed in training, which we find does not noticeably affect the results.

4 Experiments

In this section, we evaluate the proposed methods on two cross-lingual tasks: cross-lingual entity matching, and triple-wise alignment verification. We also conduct experiments on two monolingual tasks. Besides, a *case study* with knowledge alignment examples is included in the Appendix of [Chen *et al.*, 2017].

Data Sets. Experimental results on the trilingual data sets WK31 are reported in this section. WK31 contains English (En), French (Fr), and German (De) knowledge graphs under DBpedia’s `dbo:Person` domain, where a part of triples are aligned by verifying the ILLs on entities, and multi-lingual labels of the DBpedia ontology on some relations. The number of entities in each language is adjusted to obtain two data sets. For each of the three languages thereof, WK31-15k matches the number of nodes (about 15,000) with FB15k—the largest monolingual graph used by many recent works [Zhong *et al.*, 2015; Lin *et al.*, 2015; Ji *et al.*, 2015; Jia *et al.*, 2016], and the number of nodes in WK31-120k is several times larger. For both data sets, German graphs are sparser than English and French graphs. We also collect extra entity ILLs for the evaluation of cross-lingual entity matching, whose quantity is shown in Table 3. Meanwhile, we derive another trilingual data set CN31 from ConceptNet [Speer and Havasi, 2013]. Additional results on CN31 that lead to similar evaluation conclusions are reported in the Appendix of [Chen *et al.*, 2017].

4.1 Cross-lingual Entity Matching

The objective of this task is to match the same entities from different languages in KB . Due to the large candidate space, this task emphasizes more on ranking a set of candidates rather than acquiring the best answer. We perform this task on both data sets to compare five variants of MTransE.

To show the superiority of MTransE, we adapt LM, CCA, and OT (which are introduced in Section 2) to their knowledge graph equivalences.

Evaluation Protocol. Each MTransE variant is trained on a complete data set. LM and CCA are implemented by in-

Table 4: Cross-lingual entity matching result.

Data Set	WK31-15k								WK31-120k			
	En-Fr		Fr-En		En-De		De-En		En-Fr	Fr-En	En-De	De-En
Aligned Languages	Hits@10	Mean	Hits@10	Mean	Hits@10	Mean	Hits@10	Mean	Hits@10	Hits@10	Hits@10	Hits@10
LM	12.31	3621.17	10.42	3660.98	22.17	5891.13	15.21	6114.08	11.74	14.26	24.52	13.58
CCA	20.78	3094.25	19.44	3017.90	26.46	5550.89	22.30	5855.61	19.47	12.85	25.54	20.39
OT	44.97	508.39	40.92	461.18	44.47	155.47	49.24	145.47	38.91	37.19	38.85	34.21
Var ₁	51.05	470.29	46.64	436.47	48.67	146.13	50.60	167.02	38.58	36.52	42.06	47.79
Var ₂	45.25	570.72	41.74	565.38	46.27	168.33	49.00	211.94	31.88	30.84	41.22	40.39
Var ₃	38.64	587.46	36.44	464.64	50.82	125.15	52.16	151.84	38.26	36.45	50.48	52.24
Var ₄	59.24	190.26	57.48	199.64	66.25	74.62	68.53	42.31	48.66	47.43	57.56	63.49
Var ₅	59.52	191.36	57.07	204.45	60.25	99.48	66.03	54.69	45.65	47.48	64.22	67.85

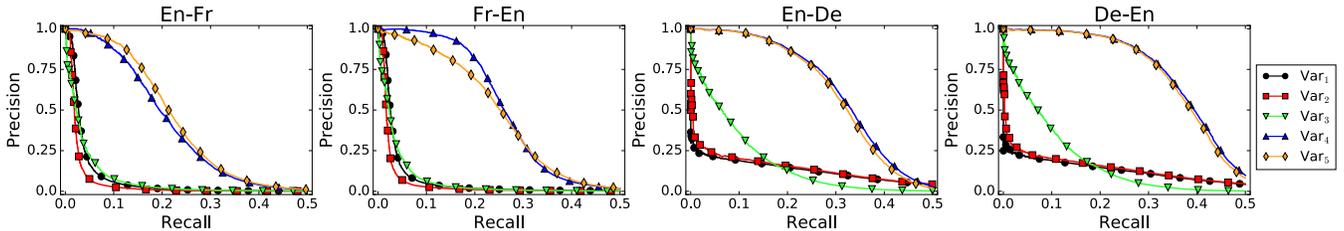


Figure 1: Precision-recall curves for cross-lingual entity matching on WK31-15k.

ducing the corresponding transformations across separately trained knowledge models on monolingual graphs, while using the alignment sets as anchors. Training OT is quite similar to MTransE, we add the process of orthogonalization to the training of the alignment model, since the regularization of vectors has already been enforced. The entity ILLs are used as ground truth for test. We take these unidirectional links between English-French and English-German, i.e., four directions in total. For each ILL (e, e') , we perform a kNN search from the cross-lingual transition point of e (i.e., $\tau(e)$) and record the rank of e' . Following the convention [Xing *et al.*, 2015; Jia *et al.*, 2016], we aggregate two metrics over all test cases, i.e., the proportion of ranks no larger than 10 *Hits@10* (in percentage), and the mean rank *Mean*. We prefer higher *Hits@10* and lower *Mean* that indicate a better outcome.

For training, we select the learning rate λ among $\{0.001, 0.01, 0.1\}$, α among $\{1, 2.5, 5, 7.5\}$, l_1 or l_2 norm in loss functions, and dimensionality k among $\{50, 75, 100, 125\}$. The best configuration on WK31-15k is $\lambda = 0.01$, $\alpha = 5$, $k = 75$, l_1 norm for Var₁, Var₂, LM, and CCA, l_2 norm for other variants and OT. While the best configuration on WK31-120k is $\lambda = 0.01$, $\alpha = 5$, $k = 100$, and l_2 norm for all models. The training on both data sets takes 400 epochs.

Results. We report *Hits@10* and *Mean* for WK31-15k, and *Hits@10* for WK31-120k, on the four involved directions of cross-lingual matching in Table 4. As expected, without jointly adapting the monolingual vector spaces with the knowledge alignment, LM and CCA are largely outperformed by the rest. While the orthogonality constraint being too strong to be enforced in these cases, OT performs at most closely to the simplest cases of MTransE. For MTransE, Var₄ and Var₅ outperform the other three variants under all settings. The fairly close results obtained by these two variants indicate that the interference caused by learning an additional relation-dedicated transformation in Var₅ is negligible to the entity-dedicated transformation. Correspondingly, we believe that the reason for Var₃ to be outperformed by Var₄ and Var₅ is that it fails to differentiate well the over-frequent cross-lingual alignment from regular relations. Therefore, the char-

acterization for cross-lingual alignment is negatively affected by the learning process for monolingual relations in a visible degree. Axis calibration appears to be unstable on this task. We hypothesize that this simple technique is affected by two factors: coherence between language-specific versions, and density of the graphs. Var₂ is always outperformed by Var₁ due to the negative effect of the calibration based on relations. We believe this is because multi-mapping relations are not so well-captured by TransE as explained in [Wang *et al.*, 2014], therefore disturb the calibration of the entire embedding spaces. Although Var₁ still outperforms Var₃ on entity matching between English and French graphs in WK31-15k, coherence somewhat drops alongside when scaling up to the larger data set so as to hinder the calibration. The German graphs are sparse, thus should have set a barrier for precisely constructing embedding vectors and hindered calibration on the other side. Therefore Var₁ still performs closely to Var₃ in the English-German task on WK31-15k and English-French task on WK31-120k, but is outperformed by Var₃ in the last setting. In general, the variants that use linear transformations are the most desired. This conclusion is supported by their promising outcome on this task, and it is also reflected in the precision-recall curves shown in Figure 1.

4.2 Triple-wise Alignment Verification

This task is to verify whether a given pair of aligned triples are truly cross-lingual counterparts. It produces a classifier that helps with verifying candidates of triple matching [Nguyen *et al.*, 2011; Rinser *et al.*, 2013].

Evaluation Protocol. We create positive cases by isolating 20% of the alignment set. Similar to [Socher *et al.*, 2013], we randomly corrupt positive cases to generate negative cases. In detail, given a pair of correctly aligned triples (T, T') , it is corrupted by (i) randomly replacing one of the six elements in the two triples with another element from the same language, or (ii) randomly substituting either T or T' with another triple from the same language. Cases (i) and (ii) respectively contribute negative cases that are as many as 100% and 50% of positive cases. We use 10-fold cross-validation on these cases to train and evaluate the classifier.

Table 5: Accuracy of TWA verification (%).

Data Set	WK31-15k		WK31-120k	
	En&Fr	En&De	En&Fr	En&De
LM	52.23	63.61	59.98	59.98
CCA	52.28	66.49	65.89	61.01
OT	93.20	87.97	88.65	85.24
Var ₁	93.25	91.24	91.27	91.35
Var ₂	90.24	86.59	89.36	86.29
Var ₃	90.38	84.24	87.99	87.04
Var ₄	94.58	95.03	93.48	93.06
Var ₅	94.90	94.95	92.63	93.66

Table 6: Results of tail prediction (*Hits@10*).

Data Set	WK31-15k		WK31-120k	
	En	Fr	En	Fr
TransE	42.19	25.06	36.78	25.38
Var ₁	40.37	23.45	39.09	25.52
Var ₂	40.80	24.77	36.02	21.13
Var ₃	40.97	22.26	35.99	19.69
Var ₄	41.03	25.46	39.64	25.59
Var ₅	41.79	25.77	38.35	24.68

Table 7: Results of relation prediction (*Hits@10*).

Data Set	WK31-15k		WK31-120k	
	En	Fr	En	Fr
TransE	61.79	62.55	60.06	65.29
Var ₁	60.18	60.73	61.75	65.46
Var ₂	54.33	62.98	61.11	61.47
Var ₃	58.32	59.44	60.14	48.06
Var ₄	63.74	64.77	60.26	67.64
Var ₅	64.79	63.71	60.77	66.86

We use a simple threshold-based classifier similar to the widely-used ones for triple classification [Socher *et al.*, 2013; Wang *et al.*, 2014; Lin *et al.*, 2015]. For a given pair of aligned triples $(T, T') = ((h, r, t), (h', r', t'))$, the dissimilarity function is defined as $f_d(T, T') = \|\tau(\mathbf{h}) - \mathbf{h}'\|_2 + \|\tau(\mathbf{r}) - \mathbf{r}'\|_2 + \|\tau(\mathbf{t}) - \mathbf{t}'\|_2$. The classifier finds a threshold σ such that $f_d < \sigma$ implies positive, otherwise negative. The value of σ is determined by maximizing the accuracy for each fold on the training set. Such a simple classification rule adequately relies on how precisely each model represents cross-lingual transitions for both entities and relations.

We carry forward the corresponding configuration from the last experiment, just to show the performance of each variant under controlled variables.

Results. Table 5 shows the mean accuracy, with a standard deviation below 0.009 in cross-validation for all settings. Thus, the results are statistically sufficient to reflect the performance of classifiers. Note that the results appear to be better than those of the previous task since this is a binary classification problem. Intuitively, the linear-transformation-based MTransE perform steadily and take the lead on all settings. We also observe that Var₅, though learns an additional relation-dedicated transformation, still performs considerably close to Var₄ (the difference is at most 0.85%). The simple Var₁ is the runner-up, and is between 1.65% and 3.79% to the optimal solutions. However the relation-dedicated calibration in Var₂ causes a notable setback (4.12%~8.44% from the optimal). The performance of Var₃ falls behind slightly more than Var₂ (4.52%~10.79% from the optimal) due to the failure in distinguishing cross-lingual alignment from regular relations. Meanwhile, we single out the accuracy on the portion of negative cases where only the relation is corrupted for English-French in WK31-15k. The five variants receive 97.73%, 93.78%, 82.34%, 98.57%, and 98.54%, respectively. The close accuracy of Var₄ and Var₅ indicates that the only transformation learnt from entities in Var₄ is enough to substitute the relation-dedicated transformation in Var₅ for discriminating relation alignment, while learning the additional transformation in Var₅ does not notably interfere the original one. However, it applies differently to axis calibration since Var₂ does not improve but actually impairs the cross-lingual transitions for relations. For the same reasons as above, LM and CCA do not match with MTransE in this experiment as well, while OT performs closely to some variants of MTransE, but is still left behind by Var₄ and Var₅.

4.3 Monolingual Tasks

The above experiments have shown the strong capability of MTransE in handling cross-lingual tasks. Now we report the results on comparing MTransE with its monolingual coun-

terpart TransE on two monolingual tasks introduced in the literature [Bordes *et al.*, 2013; 2014], namely tail prediction (predicting t given h and r) and relation prediction (predicting r given h and t), using the English and French versions of our data sets. Like previous works [Bordes *et al.*, 2013; Wang *et al.*, 2014; Jia *et al.*, 2016], for each language version, 10% triples are selected as the test set, and the remaining becomes the training set. Each MTransE variant is trained upon both language versions of the training set for the knowledge model, while the intersection between the alignment set and the training set is used for the alignment models. TransE is trained on either language version of the training set. Again, we use the configuration from the previous experiment.

Results. The results for *Hits@10* are reported in Tables 6 and 7. They imply that MTransE preserves well the characterization of monolingual knowledge. For each setting, Var₁, Var₄, and Var₅ perform at least as well as TransE, and some even outperforms TransE under certain settings. This signifies that the alignment model does not interfere much with the knowledge model in characterizing monolingual relations, but might have actually strengthened it since coherent portions of knowledge are unified by the alignment model. Since such coherence is currently not measured, this question is left as a future work. The other question that deserves further attention is, how other knowledge models involving relation-specific entity transformations [Wang *et al.*, 2014; Lin *et al.*, 2015; Ji *et al.*, 2015; Jia *et al.*, 2016; Nguyen *et al.*, 2016] may influence monolingual and cross-lingual tasks.

5 Conclusion and Future Work

At the best of our knowledge, this paper is the first work that generalizes knowledge graph embeddings to the multilingual scenario. Our model MTransE characterizes monolingual relations and compares three different techniques to learn cross-lingual alignment for entities and relations. Extensive experiments on the tasks of cross-lingual entity matching and triple alignment verification show that the linear-transformation-technique is the best among the three. Moreover, MTransE preserves the key properties of monolingual knowledge graph embeddings on monolingual tasks.

The results here are very encouraging, but we also point out opportunities for further work and improvements. In particular, we should explore how to substitute the simple loss function of the knowledge model used in MTransE with more advanced ones involving relation-specific entity transformations. More sophisticated tasks of cross-lingual triple completion may also be conducted. Combining MTransE with multilingual word embeddings [Xing *et al.*, 2015] is another meaningful direction since it will provide a useful tool to extract new relations from multilingual text corpora.

References

- [Bond and Foster, 2013] Francis Bond and Ryan Foster. Linking and extending an open multilingual Wordnet. In *ACL*, pages 1352–1362, 2013.
- [Bordes *et al.*, 2011] Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning structured embeddings of knowledge bases. In *AAAI*, 2011.
- [Bordes *et al.*, 2012] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. Joint learning of words and meaning representations for open-text semantic parsing. In *AISTATS*, pages 127–135, 2012.
- [Bordes *et al.*, 2013] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795, 2013.
- [Bordes *et al.*, 2014] Antoine Bordes, Jason Weston, and Nicolas Usunier. Open question answering with weakly supervised embedding models. In *ECML-PKDD*, pages 165–180, 2014.
- [Chen *et al.*, 2017] Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. Multi-lingual knowledge graph embeddings for cross-lingual knowledge alignment. *arXiv preprint arXiv:1611.03954*, 2017.
- [Collobert and Weston, 2008] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *ICML*, pages 160–167, 2008.
- [Culotta and Sorensen, 2004] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In *ACL*, page 423, 2004.
- [Faruqui and Dyer, 2014] Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. *EACL*, 2014.
- [Jenatton *et al.*, 2012] Rodolphe Jenatton, Nicolas L Roux, Antoine Bordes, and Guillaume R Obozinski. A latent factor model for highly multi-relational data. In *NIPS*, 2012.
- [Ji *et al.*, 2015] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. Knowledge graph embedding via dynamic mapping matrix. In *ACL*, pages 687–696, 2015.
- [Jia *et al.*, 2016] Yantao Jia, Yuanzhuo Wang, Hailun Lin, Xiaolong Jin, and Xueqi Cheng. Locally adaptive translation for knowledge graph embedding. In *AAAI*, 2016.
- [Lehmann *et al.*, 2015] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, et al. Dbpedia—a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195, 2015.
- [Lin *et al.*, 2015] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, 2015.
- [Mahdisoltani *et al.*, 2015] Farzaneh Mahdisoltani, Joanna Biega, Fabian Suchanek, et al. Yago3: A knowledge base from multilingual Wikipedias. In *CIDR*, 2015.
- [Mikolov *et al.*, 2013] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. *arXiv*, 2013.
- [Nguyen *et al.*, 2011] Thanh Nguyen, Viviane Moreira, Huong Nguyen, Hoa Nguyen, and Juliana Freire. Multilingual schema matching for Wikipedia infoboxes. *PVLDB*, 5(2):133–144, 2011.
- [Nguyen *et al.*, 2016] Dat Quoc Nguyen, Kairit Sirts, Lizhen Qu, and Mark Johnson. Stranse: a novel embedding model of entities and relationships in knowledge bases. In *NAA-CL HLT*, pages 460–466, 2016.
- [Nickel *et al.*, 2016] Maximilian Nickel, Lorenzo Rosasco, Tomaso Poggio, et al. Holographic embeddings of knowledge graphs. In *AAAI*, 2016.
- [Rinser *et al.*, 2013] Daniel Rinser, Dustin Lange, Felix Naumann, and Gerhard Weikum. Cross-lingual entity matching and infobox alignment in Wikipedia. *Information Systems*, 38(6):887–907, 2013.
- [Saxe *et al.*, 2014] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *ICLR*, 2014.
- [Socher *et al.*, 2013] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *NIPS*, pages 926–934, 2013.
- [Speer and Havasi, 2013] Robert Speer and Catherine Havasi. Conceptnet 5: A large semantic network for relational knowledge. *The People’s Web Meets NLP*, pages 161–176, 2013.
- [Suchanek *et al.*, 2011] Fabian M Suchanek, Serge Abiteboul, Pierre Senellart, and Tom Mitchell. Paris: Probabilistic alignment of relations, instances, and schema. *PVLDB*, 5(3):157–168, 2011.
- [Sun *et al.*, 2011] Ang Sun, Ralph Grishman, and Satoshi Sekine. Semi-supervised relation extraction with large-scale word clustering. In *ACL*, pages 521–529, 2011.
- [Vrandečić, 2012] Denny Vrandečić. Wikidata: A new platform for collaborative data collection. In *WWW*, 2012.
- [Wang *et al.*, 2014] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, 2014.
- [Weston *et al.*, 2013] Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. Connecting language and knowledge bases with embedding models for relation extraction. In *EMNLP*, pages 1366–1371, 2013.
- [Wikipedia, 2017] Wikipedia, 2017. <https://www.wikipedia.org/>.
- [Wilson and Martinez, 2003] D Randall Wilson and Tony R Martinez. The general inefficiency of batch training for gradient descent learning. *Neural Networks*, 16(10), 2003.
- [Xing *et al.*, 2015] Chao Xing, Dong Wang, Chao Liu, and Yiye Lin. Normalized word embedding and orthogonal transform for bilingual word translation. In *NAACL HLT*, pages 1006–1011, 2015.
- [Yang *et al.*, 2015a] Cheng Yang, Zhiyuan Liu, Deli Zhao, Maosong Sun, and Edward Chang. Network representation learning with rich text information. In *IJCAI*, 2015.
- [Yang *et al.*, 2015b] Yang Yang, Yizhou Sun, Jie Tang, Bo Ma, and Juanzi Li. Entity matching across heterogeneous sources. In *KDD*, pages 1395–1404, 2015.
- [Zhong *et al.*, 2015] Huaping Zhong, Jianwen Zhang, Zhen Wang, Hai Wan, and Zheng Chen. Aligning knowledge and text embeddings by entity descriptions. In *EMNLP*, pages 267–272, 2015.
- [Zhou *et al.*, 2005] Guodong Zhou, Jian Su, Jie Zhang, and Min Zhang. Exploring various knowledge in relation extraction. In *ACL*, pages 427–434, 2005.