Multi-lingual Knowledge Graph Embeddings for Cross-lingual Knowledge Alignment

Muhao Chen¹, Yingtao Tian², Mohan Yang¹, Carlo Zaniolo¹
{muhaochen, yang, zaniolo}@cs.ucla.edu; yittian@cs.stonybrook.edu
Department of Computer Science, UCLA¹
Department of Computer Science, Stony Brook University²

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 translations, 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 \(h\) and \(t\), respectively. A function \(f_r(h, t)\) is used to measure the plausibility of \((h, r, t)\), which also implies the transformation \(r\) that characterizes \(r\).

Exemplarily, the translation-based model TransE [Bordes et al., 2013] uses the loss function \(f_r(h, t) = \|h + r - t\|_l\), where \(r\) 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 transitions.\footnote{We use the word \textit{transition} here to differentiate from the relational translations among entities in translation-based methods.} 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 \textit{MTransE}, that learns the multilingual knowledge graph structure using a combination of two component models, namely \textit{knowledge model} and \textit{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 \(h + r \approx t\), where \(h\) and \(t\) 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 \cite{Bordes+al-13} sets \(h\) and \(t\) as the original \(h\) and \(t\), 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 \(h\) and \(t\), including projections on relation-specific hyperplanes in TransH \cite{Wang+al-14}, linear transformations to heterogeneous relation spaces in TransR \cite{Lin+al-15}, and other forms \cite{Jia+al-15,Nguyen+al-16}. 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 \cite{Westen+al-13, Zhong+al-15}.

In addition to these, there are non-translation-based methods. Some of those including UM \cite{Bordes+al-15}, SE \cite{Bordes+al-12}, Bilinear \cite{Jenatton+al-12}, and HolE \cite{Nickel+al-16}, do not explicitly represent relation embeddings. Others including neural-based models SLM \cite{Collobert-08} and NTN \cite{Socher+al-13}, and random-walk-based model TADW \cite{Yang+al-15}, 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. Of those can be extended to multilingual knowledge graphs, such as LM \cite{Mikolov+al-13} and CCA \cite{Faruqui+al-13} which induce offline transitions among pre-trained monolingual embeddings in forms of linear transformations and canonical component 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 \cite{Xing+al-15} 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 \cite{Vrandecic+12} and DBpedia \cite{Lehmann+al-15} rely on crowdsourcing to create LILS and relation alignment. YAGO \cite{Mahdisoltani+al-15} 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 \cite{Nguyen+al-11, Suchanek+al-11, Rinser+al-13} to entity descriptions \cite{Yang+al-15}, which being unavailable to many knowledge bases such as YAGO, WordNet, and ConceptNet \cite{Speer+Havasi-13}. 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.

3.1 Multilingual Knowledge Graphs

In a knowledge base $KB$, we use $L$ to denote the set of languages, and $L^2$ to denote the 2-combination of $L$ (i.e., the set of unordered language pairs). For a language $L \in L$, $G_L$ 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 $h$, $r$, $t$ respectively represent the embedding vectors of head $h$, relation $r$, and tail $t$. For a language pair $(L_1, L_2) \in L^2$, $\delta(L_1, L_2)$ 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 (State of California, capital city, Sacramento, (État de Californie, capitale, Sacramento)) $\in \delta$(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 $L^2$ 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 $(L_i, L_j) \in L^2$ as an example to describe how we define each component of a model.

3.2 Knowledge Model

For each language $L \in L$, a dedicated $k$-dimensional embedding space $\mathbb{R}^k_L$ 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 as below:

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

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:

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

Thereof, 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.

$$S_{a1} = \|h - h'| + \|t - t'||$$

$S_{a1}$ regulates that correctly aligned multilingual expressions of the same entity tend to have close embedding vectors. Thus by minimizing the loss function that involves $S_{a1}$, 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.

$$S_{a2} = \|h - h'| + \|r - r'| + \|t - t'||$$

$S_{a2}$ overlays the penalty of relation alignment to $S_{a1}$, 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 $S_{a3}$ as below is derived.

$$S_{a3} = \|h + v_{ij}^e - h'| + \|r + v_{ij}^r - r'| + \|t + v_{ij}^t - t'||$$

Thereof $v_{ij}^e$ and $v_{ij}^r$ are respectively deployed as the entity-dedicated and relation-dedicated translation vectors between $L_i$ and $L_j$, such that we have $e + v_{ij}^e \approx e'$ for embedding vectors $e$, $e'$ of the same entity $e$ expressed in both languages, and $r + v_{ij}^r \approx r'$ 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 $v_{ij}^e = -v_{ji}^e$ and $v_{ij}^r = -v_{ji}^r$ 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. $S_{a4}$ as below learns a $k \times k$ square matrix $M_{ij}^e$ as a linear transformation on entity vectors from $L_i$ to $L_j$, given $k$ as the dimensionality of the embedding spaces.

$$S_{a4} = \|M_{ij}^e h - h'| + \|M_{ij}^t t - t'||$$


Table 1: Summary of model variants.

| Var | Model Complexity | Cross-lingual Transition | Search Complexity |
|-----|-------------------|--------------------------|-------------------|
| Var1 | O(n, k, l) + n, k, l | \(\tau_j(e) = e\) | O(n, k) |
| Var2 | O(n, k, l) + n, k, l | \(\tau_j(e) = r\) | O(n, k) |
| Var3 | O(k, l) + n, k, l | \(\tau_j(e) = e + v_j\) | O(n, k) |
| Var4 | O(k, l) + n, k, l | \(\tau_j(e) = r\sum_k^l\tau_j(e)\) | O(n, k) |
| Var5 | O(k, l) + n, k, l | \(\tau_j(e) = r\sum_k^l\tau_j(e)\) | O(n, k) |

Notation: \(e\) and \(r\) respectively the vectors of an entity \(e\) and a relation \(r\), \(k\) is the dimension of the embedding spaces, \(l\) is the cardinality of \(L\), \(n_e\) and \(n_r\) are respectively the number of entities and the number of relations, where \(n_e \gg n_r\), \(S_{a5}\) additionally brings in a second linear transformation \(M_{ij}^r\) for relation vectors, which is of the same shape as \(M_{ij}^e\). The use of a different matrix is again due to different redundancy of entities and relations.

\[
S_{a5} = \|M_{ij}^e h - h'\| + \|M_{ij}^r r - r'\| + \|M_{ij}^e t - t'\|
\]

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 \(M_{ij}^{-1} = M_{ij}^e\) and \(M_{ij}^{-1} = M_{ij}^r\). Thus the transition in the reverse 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 \(\alpha\) 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\(\_e\) as the variant of MTransE that adopts the \(k\)-th alignment model which employs \(S_{nk}\). 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 \(\tau_j\) that maps a cross-lingual transition of a vector from \(L_i\) to \(L_j\), or simply \(\tau\) 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 \(L^2\). 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 \(\theta\) by setting \(\theta = \theta - \lambda \nabla J\), where \(\lambda\) is the learning rate. Instead of directly updating \(J\), our implementation optimizes \(S_K\) and \(\alpha S_A\) 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; Bordes et al., 2014; Jernatton et al., 2012] and has two important effects: (i) it helps avoid the case 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 Var\(\_e\) and Var\(\_r\).

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 [Anonym. 2017].

Data Sets. Experimental results on the trilingual data sets WK3l are reported in this section. WK3l 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. Thereof, for each of the three languages, WK3l-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 WK3l-120k is several times larger. For both data sets, German graphs are sparser than English and French graphs. We also collect extra ILLs that are not covered by the alignment sets for the evaluation of cross-lingual entity matching, whose quantity is shown in Table 2. Meanwhile, we derive another trilingual data set CN3l from ConceptNet [Speer and Havasi, 2013]. Additional results on CN3l that lead to similar evaluation conclusions are reported in the Appendix of [Anonym. 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.
Table 4: Cross-lingual entity matching result.

| Data Set       | Aligned Languages | WK3l-13k | WK3l-120k |
|----------------|-------------------|----------|-----------|
|                | En-Fr             | Fr-En    | En-De     | De-En     |
|                | Hits@10 Mean      | Hits@10  | Hits@10   | Hits@10   |
| LM             | 12.31 362.13      | 19.44 3017.9 | 22.17 5507.13 | 15.21 6114.08 |
| CCA            | 20.78 3094.25     | 26.46 5550.89 | 22.30 5585.61 | 19.47 128.54 |
| OT             | 44.97 508.39      | 44.47 155.47 | 49.24 145.47 | 38.91 37.19  |
| Var1           | 51.05 470.29      | 48.67 146.13 | 50.60 167.02 | 38.58 36.52  |
| Var2           | 45.25 570.72      | 47.79 168.33 | 49.00 211.94 | 31.88 30.84  |
| Var3           | 38.64 587.46      | 50.82 125.15 | 51.26 151.84 | 38.26 36.45  |
| Var4           | 59.24 190.26      | 66.25 74.62 | 68.53 86.33 | 48.66 47.43  |
| Var5           | 59.52 191.36      | 60.25 99.48 | 66.03 95.79 | 45.65 47.48  |

Figure 1: Precision-recall curves for cross-lingual entity matching on WK3l-15k.

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 inducing 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 extra ILLs are used as ground truth for test. We take these unidirectional links closely to the simplest cases of MTransE. For MTransE, Var 4 is always outperformed by Var3 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., 2015, Jia et al., 2016], therefore disturb the calibration of the entire embedding spaces. Although Var3 still precedes Var2 on entity matching between English and French graphs in WK3l-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 WK3l-15k and English-French task on WK3l-120k, but is preceded by Var2 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.

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 $f_d(T, T') = \|\mathbf{r}(h) - \mathbf{r}(h')\|_2 + \|\mathbf{r}(t) - \mathbf{t'}\|_2$. The classifier finds a threshold $\sigma$ such that $f_d < \sigma$ implies positive, otherwise negative. The value of $\sigma$ 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 Var5 is enough to substitute the relation-dedicated transformation in Var5 for discriminating relation alignment, while learning the additional transformation in Var3 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 counterpart TransE on two monolingual tasks introduced in the literature [Bordes et al., 2013; Bordes et al., 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

[Anonymous, 2017] Anonymous, 2017. https://dl.dropboxusercontent.com/u/49886343/multigraph.pdf

[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.

[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.

[Cutollo and Sorensen, 2004] Aron Cutollo 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.

[Janott et al., 2012] Rodolphe Janott, 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, Haijun 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 NAACL 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.
6 Appendix

6.1 Examples of Knowledge Alignment

We have already shown the effectiveness of MTransE in aligning cross-lingual knowledge, especially the linear-transformation-based variants Var4 and Var5. Now we discuss several examples to reveal insights on how our methods may be used in cross-lingual knowledge augmentation.

Table 8: Examples of cross-lingual entity matching.

| Entity | German | French |
|--------|--------|--------|
| Paris  | Frankreich, Los Angeles, Santa Monica, Californie | Paris, Los Angeles, Palm Springs, Santa Monica |
| California | | 
| rock music | rockmusik, soul, death metal, dance-pop | 

Table 9: Examples of cross-lingual relation matching.

| Relation | Target |
|----------|--------|
| capital  | Frankreich, pays d'accueil, lieu de naissance |
| nationality | nation, nationalität, nationalité, nacionalidad |
| language  | langue, réalisations, lieu de naissance, nationalité |
| nickname  | surnom, alternativnamen |

Table 10: Examples of cross-lingual entity matching.

| Query | Target |
|-------|--------|
| (Adam Lambert, genre, ?) | musique indépendante, musique alternative, rock, glam rock |
| (Ronaldinho, position, ?) | milieu offensif, attaquant, quarterback, latéral gauche |
| (Italy, ?, Rome) | capitale, plus grande ville, chef-lieu, garnison |
| (Barack Obama, ?, George Bush) | chef-lieu, capitale, premier ministre, président du conseil |
| (? instrument, guitar) | guitare, guitar, chitarra, gitaar |

We start with the search of cross-lingual counterparts of entities and relations. We choose an entity (or relation) in English and then show the nearest candidates in French and German, respectively. These candidates are listed by decreasing values of the Euclidean distance between their vectors in the target language space and the result point of cross-lingual transition. Several examples are shown in Table 8 and Table 9. In all tables of this subsection, we mark the exact answers as boldfaced, and the conceptually close ones as italic. For example, in Table 8 besides boldfacing the exactly correct answers for Barack Obama and Paris, we consider those who have also been U.S. presidents as conceptually close to Paris. Also, in Table 9, those French and German relations that have the meaning of settlements of significance are considered as conceptually close to capital.

We then move on to the more complicated cross-lingual triple completion task. We construct queries by replacing one element in an English triple with a question mark, for which we seek answers in another language. Our methods need to transfer the remaining elements to the space of the target language and pick the best answer for the missing element. Table 10 shows some query answers. It is noteworthy that the basic queries are already useful for aided cross-lingual augmentation of knowledge. However, developing a joint model to support complex queries on multilingual knowledge graphs based on MTransE generated features appears to be a promising future work to support Q&A on multilingual knowledge bases.

Figure 2 shows the PCA projection of the same six English entities in their original English space and in French space after transformation. We can observe that the vectors of English entities show certain structures, where the U.S. cities are grouped together and other countries’ cities are well separated. After transformation into French space, these English entities not only keep their original spatial emergence, but also are close to their corresponding entities in French. This illustrates the transformation preserves mono-lingual structure and also it is able to capture cross-lingual information. We believe this example illustrates the good performance and also it is able to capture cross-lingual information.

6.2 Additional Experimental Results

We derive another data set CN3l from the MIT ConceptNet to evaluate MTransE, whose statistics are shown in Table 11. Though a smaller data set than WK3l-15k, knowledge graphs in CN3l are highly sparse. Thereof, each language version of CN3l contains around 7,500 nodes and less than 41 types of relations. The alignment sets are created based on the relation TranslationOff of the ConceptNet. In this section we report the results of the two cross-lingual tasks on the CN3l data set for MTransE as well as all baselines. Basically, these results lead to similar conclusions as we have on WK3l.

Evaluation Protocol. The metrics and evaluation procedures are the same as those on WK3l. We select λ among \{0.001, 0.01, 0.05\}, α among \{1, 2.5, 5, 7.5\}, l1 or l2 norm in loss functions, and dimensionality k among \{25, 50, 75\}. Optimal parameters are configured as λ = 0.001, α = 2.5, k = 50, and l1 norm for all models. To perform the evaluation under controlled variables, we again use one configuration on each model respectively in the two experiments. Since the data set is smaller, training is limited to 200 epochs.

Results of Cross-lingual Entity Matching. The results are reported in Table 12. For the baselines, LM and CCA are again left far behind for being disjointedly trained. OT, however, takes the position ahead of Var1. This is likely because the knowledge graphs in CN3l are highly sparse, therefore fewer interference of monolingual relations among entities makes the orthogonality constraint easier to fulfill. Even so, in all settings, OT is still largely outperformed by Var4 and Var5, which receives amazingly good results, thus steadily being the optimal solutions. Interestingly, Var3 now ranks right behind the linear-transformation-based variants in most.

| Type of triples | En-Fr | Fr-En | En-De | De-En |
|----------------|-------|-------|-------|-------|
| Number of triples | 47,696 | 18,624 | 25,560 | 3,668 |
| Type of ILLs | En-Fr | Fr-En | En-De | De-En |
| Number of ILLs | 2,134 | 2,146 | 3,485 | 3,813 |
settings. This is quite reasonable because the cross-lingual transitions, which are regarded as a type of relation by \text{Var}_3 in the graphs, are now way less frequent in CN3l than they are in the much denser and more heterogeneous WK3l. Thus, this explains why it performs better than \text{Var}_1. For the same reason as we stated in Section 4.1, \text{Var}_2 is placed at last of the five MTransE variants in matching cross-lingual entities.

**Results of Triple-wise Alignment Verification.** The results shown in Table 13 reflect the same conclusions of the experiment performed on WK3l that, the linear-transformation-based variants takes the lead ahead of the rest MTransE variants and the baselines. While \text{Var}_1, despite being the simplest, takes the second place with a satisfying accuracy in triple-wise alignment verification as well. The relation-dedicated calibration still causes interference in the optimization process of \text{Var}_2, therefore leads to a 4%~9% drop of accuracy from \text{Var}_1. \text{Var}_3 performs slightly better than \text{Var}_2. On triple-wise alignment verification on CN3l, we receive exactly the same placement for evaluating the five MTransE variants. Meanwhile, for the baselines, OT is slightly worse than \text{Var}_1, CCA also receives acceptable accuracy which is however worse than all MTransE variants, while the accuracy of LM is just slightly better than random guessing.

Above all, the results in CN3l indicates that MTransE also works promisingly on very sparse multilingual graphs, while the linear-transformation-based variants are the best representation techniques.