An Accurate Unsupervised Method for Joint Entity Alignment and Dangling Entity Detection

Shengxuan Luo[1,2] **Sheng Yu**[1,2]
[1]Center for Statistical Science, Tsinghua University
[2]Department of Industrial Engineering, Tsinghua University
luosx18@mails.tsinghua.edu.cn
syu@tsinghua.edu.cn

Abstract

Knowledge graph integration typically suffers from the widely existing dangling entities that cannot find alignment cross knowledge graphs (KGs). The dangling entity set is unavailable in most real-world scenarios, and manually mining the entity pairs that consist of entities with the same meaning is labor-consuming. In this paper, we propose a novel accurate **U**nsupervised method for joint **E**ntity alignment (EA) and **D**angling entity detection (DED), called UED. The UED mines the literal semantic information to generate pseudo entity pairs and globally guided alignment information for EA and then utilizes the EA results to assist the DED. We construct a medical cross-lingual knowledge graph dataset, MedED, providing data for both the EA and DED tasks. Extensive experiments demonstrate that in the EA task, UED achieves EA results comparable to those of state-of-the-art supervised EA baselines and outperforms the current state-of-the-art EA methods by combining supervised EA data. For the DED task, UED obtains high-quality results without supervision.

1 Introduction

Entity alignment (EA) that aligns the equivalent entities in different knowledge graphs (KGs) is a fundamental technique for knowledge graph integration. A typical application of EA is constructing a large-scale KG by integrating different KGs to facilitate various downstream tasks such as question answering (Savenkov and Agichtein, 2016; Yu et al., 2017; Jin et al., 2022), recommendation (Cao et al., 2019), and search engines (Xiong et al., 2017). The existing embedding-based EA methods align each entity to its closest counterpart cross KGs according to entity embeddings. In recent years, they have emerged as the dominant EA solutions due to their effectiveness and strong ability to utilize information such as entity name strings, entity description, attributes, and graph structure. These EA methods (Chen et al., 2017; Sun et al., 2018; Wang et al., 2018; Zhu et al., 2021a; Liu et al., 2021; Lin et al., 2021) are built upon the assumption that there exists a counterpart in the target KG for any source entity (Sun et al., 2021). Therefore, ideally, their performances are assessed by only considering the entities in the set of testing entity pairs.

In the real-world scenario, four facts should be considered when aligning KGs: (1) The entities that do not have counterparts in another KG are ubiquitous. These entities are referred to as dangling entities, following Sun et al. (2021). Therefore, it is necessary to identify the dangling entities and then align the remaining matchable entities to their counterparts. The widely used approach of integrating KGs according to the cross KG similarity between entities loses sight of identifying dangling entities. (2) Dangling entity sets are not labeled in most cases, while some entity pairs are relatively available but labor-consuming. For example, we can preliminarily obtain pseudo entity pairs with high similarity according to extra information to align entities and then manually extract the correct pairs. The extra information could be cross KG links or literal semantic information from machine translation or word embeddings. However, identifying a dangling entity requires manual comparisons between an entity and all entities in the target KG, which is tedious and almost impossible for large KGs. Dangling entity detection (DED) methods need to avoid reliance on supervision. (3) Literal semantic information has an essential impact on EA. As shown in previous works (Wu et al., 2019; Nguyen et al., 2020; Zhu et al., 2021b), competitive EA results can be achieved by translating entity names to the same language and calculating the vector representation from GloVe (Pennington et al., 2014), suggesting that it is possible to get rid of manually annotated entity pairs by automatically mining literal semantic information. (4) Align-

Findings of the Association for Computational Linguistics: ACL 2022, pages 2330 - 2339
May 22-27, 2022 ©2022 Association for Computational Linguistics

ments are associated with each other. Traditional EA methods align entities in the local alignment way by calculating the cross KGs similarity of entities and selecting the most similar entity as EA results. The local alignment neglects the association between alignment and suffers from conflicting many-to-one and many-to-many alignments.

Considering the above facts, we propose UED, an accurate **U**nsupervised method for joint **E**A and **D**ED. For EA, to automatically mine the literal semantic information, we generate pseudo entity pairs for the align loss and design a semantic-based globally guided loss to guide the alignment for all entities, not only for those in entity pairs. For DED, since verifying the dangling entity has to check all the entities in the target KG and the dangling entity set is unavailable, we add empty entities into two KGs and transfer the EA and DED tasks into a modified global optimal transport problem (OTP) to identify dangling entities relying on pseudo entity pairs only. We propose a simple but effective way to reduce the complexity of OTP. Our experiments show that the dangling entity identification mechanism also enhances the EA performance.

There are several traditional EA datasets widely used in the EA task. Nevertheless, neither dataset provides a dangling test set for DED. As mentioned above, identifying dangling entities is crucial in real-world knowledge graph integration. To demonstrate the effectiveness of our method and incentivize future studies, we construct a cross-lingual medical knowledge graph dataset with EA task and DED task, called MedED, based on the Unified Medical Language System (UMLS) (Lindberg et al., 1993).

We summarize the main contributions as follows:

- We construct a cross-lingual knowledge graph dataset to demonstrate the effect of our designs and support future studies on EA and DED.
- We propose UED, a unified unsupervised method for both EA and DED, which gets rid of supervision in both tasks and fits the real-world scenario when aligning KGs. UED mines the literal semantic information for EA and then utilizes the EA results on pseudo entity pairs to generate high-quality DED results and consequently facilitates the performance of EA.
- We conduct comprehensive experiments on both MedED and DBP15K. In the EA task, UED achieves comparable results with state-of-the-art supervised baselines, and the supervised version of UED outperforms the current state-of-the-art methods.

The source code of UED is publicly available at https://github.com/luosx18/UED.

2 Related Work

Embedding-based Entity Alignment

Embedding-based entity alignment methods build upon knowledge embedding models, which have been developing rapidly in recent years and aim to encode KGs into low-dimensional vector space. The mainstream embedding-based EA methods adopt models such as TransE (Bordes et al., 2013), GCN (Kipf and Welling, 2016), GAT (Veličković et al., 2017), and the other variants (Sun et al., 2017; Zhu et al., 2021b), to represent entities of different KGs in vector space. Then they find equivalent entity pairs between KGs in the local alignment way.

The critical point of these EA methods is to include more semantic information in KGs accurately and effectively. The semantic information comprises graph structure, attributes, and literal information, but not all KGs contain all information mentioned above. All embedding-based EA methods adopt graph structures (Chen et al., 2017), while some methods utilize attributes (Sun et al., 2017; Trisedya et al., 2019) or literal information (Xu et al., 2019; Wu et al., 2019; Zhu et al., 2021a). To alleviate the insufficiency of training data, some studies attempt to leverage bootstrapping, iterative training techniques, and self-supervised learning to enrich the training entity pairs with pseudo pairs (Sun et al., 2018; Mao et al., 2020; Liu et al., 2021). The proposed method utilizes literal semantic information to generate alignment guidance for all entities in KGs without supervision and is compatible with all graph embedding models mentioned above.

Global Entity Alignment

Local alignment ignores the fact that alignments are associated with each other, resulting in incorrect alignments and illegal many-to-one and many-to-many alignments (Xu et al., 2020; Zeng et al., 2020). Global EA methods that consider all alignments together have been proposed to mitigate

these issues but require relatively good quality local EA to avoid the accumulation of incorrect alignments. Unfortunately, according to the Hungarian algorithm (Kuhn, 1955), the complexity of finding the best alignment between two KGs of n entities is $O(n^4)$. The existing approximate global alignment methods, CEA (Zeng et al., 2020) and GM-EHD-JEA (Xu et al., 2020), reduce the complexity with extra constraints. The CEA requires the entity pairs to be stable matches and uses the deferred acceptance algorithm (DAA) to find the alignments. The GM-EHD-JEA decomposes the entire search space into many isolated subspaces and consequently restricts the cross-subspace alignment.

Dangling Entity Detection

Several recent studies emphasize the problem of dangling entities in EA tasks. Zhao et al. (2020) and Zeng et al. (2021) introduce threshold-based methods to identify dangling entities according to the distance between a source entity and its closest target entity. These two methods identify dangling entities to improve EA behavior. Sun et al. (2021) also studied the performance of DED in the supervised setting by using the dangling training set to train the classification model or marginal ranking model.

Our method transfers the global EA and the DED into a modified unified optimal transport problem and consequently relieves the constraints on global EA, utilizes the association between alignment, and does not rely on dangling entity labels.

3 UED Framework

In this section, we first briefly describe the tasks of EA and DED and then elucidate our unified unsupervised approach to solve EA along with DED. An overview of our method is depicted in Figure 1.

3.1 Task Definition

Formally, a KG is denoted as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}\}$, where $\mathcal{E} = \mathcal{D} \cup \mathcal{A}$ is the disjoint union of dangling set $\mathcal{D}$ and matchable set $\mathcal{A}$. $\mathcal{R}$ and $\mathcal{T}$ denote the set of relations and triples, respectively. For two KGs, $\mathcal{G}_1$ and $\mathcal{G}_2$, the DED task aims to find $\mathcal{D}_1$ and $\mathcal{D}_2$, while the EA task aims to find the entity pairs between the remaining set, $\mathcal{A}_1$ and $\mathcal{A}_2$.

3.2 Pseudo Entity Pairs

Manually generating entity pairs to train the embedding base EA model is labor-consuming. We automatically generate pseudo entity pairs for model

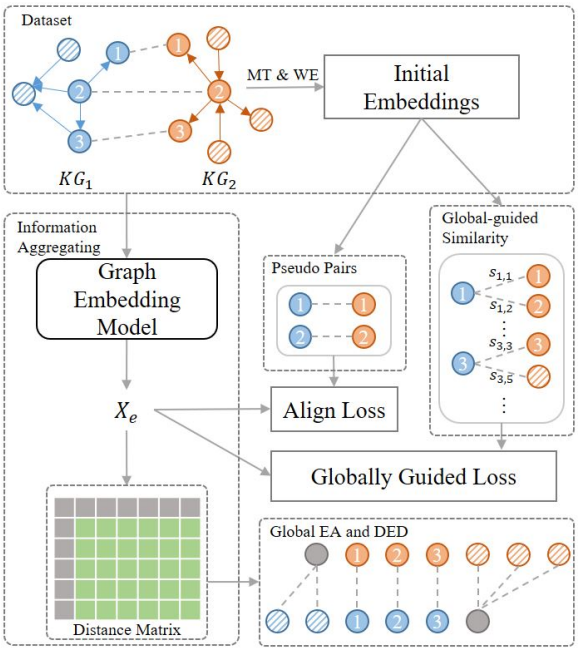

Figure 1: The Framework of UED. The rounded rectangles with dashed line denote the main modules. The circles with a number are matchable entities, and the circles with slash denote dangling entities. The gray circles are the empty entities and the gray rectangles in distance matrix denotes distance between empty entity to other entities. MT and WE refer to machine translation and word embeddings.

training, relying only on machine translation and word embeddings.

In our approach, we utilize GloVe (Pennington et al., 2014) word embeddings to generate the mean word vector v_i for entity e_i based on the entity name. Then the initial similarity between $e_i \in \mathcal{G}_1$ and $e_j \in \mathcal{G}_2$ is defined as the cosine similarity $s_{ij} = \cos(v_i, v_j)$. The set of pseudo entity pairs consists of entity pairs with high similarity. Specifically, we define a threshold $\varepsilon < 1$. If s_{ij} satisfies:

$$\begin{aligned} &s_{ij} > \varepsilon, \\ &s_{ik} \leq \varepsilon, \forall k \neq j, \\ &s_{lj} \leq \varepsilon, \forall l \neq i, \end{aligned} \tag{1}$$

then pair (e_i, e_j) is added to the pseudo entity pairs set $\mathcal{P}$. For cross-lingual KGs, we translate entity names using machine translation before applying the word embeddings.

3.3 Information Aggregating

Our method is compatible with all graph embedding models. In this paper, we follow the widespread setting to use relation triples as graph structure information and entity names as literal

information (Xu et al., 2019; Wu et al., 2019; Mao et al., 2020; Nguyen et al., 2020; Zhu et al., 2021a). We use a graph embedding model to aggregate the initial embeddings and relation triples to generate enhanced entity embeddings, X_e.

Unlike previous works (Xu et al., 2019; Mao et al., 2020; Nguyen et al., 2020; Zhu et al., 2021a), we use pseudo entity pairs to train the graph embedding model instead of training entity pairs. Denoting X_{e_i} as the output embeddings of entity e_i after the graph embedding model, we modify the hinge loss with the pseudo entity pairs, denoted as align loss:

$$\mathcal{L}_a = \sum_{(e_i,e_j)\in\mathcal{P}} \sum_{(e_i',e_j')\in\mathcal{P}'(e_i,e_j)} \max\left(d\left(X_{e_i}, X_{e_j}\right) - d\left(X_{e_i'}, X_{e_j'}\right) + \lambda, 0\right), \tag{2}$$

where λ is the margin, $\mathcal{P}'(e_i, e_j)$ is the set of negative samples for (e_i, e_j) by replacing e_i or e_j with their neighbors, and $d(\cdot,\cdot)$ is the Manhattan distance following previous works (Wu et al., 2019; Zhu et al., 2021a).

3.4 Globally Guided Similarity and Loss

The align loss does not make full use of literal semantic information since the initial similarity s_{ij} contains entity alignment information for entities not in $\mathcal{P}$. In addition, training an EA model with the align loss may mislead the model to pay too much attention to the entities in $\mathcal{P}$. Therefore, we regard entities in the target KG as anchors to guide the EA training for all source entities. Our assumption is that the counterpart of an entity is more likely to occur among entities whose initial embeddings are more similar. Specifically, we propose a globally guided loss:

$$\mathcal{L}_g = \sum_{(e_i,e_j)\in\mathcal{Q}} s_{ij} \sum_{(e_i',e_j')\in\mathcal{Q}'(e_i,e_j)} \max\left(d\left(X_{e_i}, X_{e_j}\right) - d\left(X_{e_i'}, X_{e_j'}\right) + \lambda, 0\right), \tag{3}$$

where $\mathcal{Q}$ consists of all (e_i, e_j) satisfying e_j is one of the top k similar entities of e_i according to the initial semantic similarity $\{s_{ij}, \forall j\}$, and k is a hyperparameter. The construction of $\mathcal{Q}'$ is similar to $\mathcal{P}'$. According to our experiments, s_{ij} is a necessary value that refers to the weight of (e_i, e_j) in $\mathcal{L}_g$ to improve model performance. To gradually reduce the impact of entities in $\mathcal{Q}$, we design a mechanism to decrease the weight of the globally guided loss. The final loss is

$$\mathcal{L} = \mathcal{L}_a + w(t)\mathcal{L}_g, \tag{4}$$

where t is the training step, and $w(t)$ decreases linearly to 0 as t increases.

3.5 Global EA and DED

Given two KGs comprising n and m entities, we define a distance matrix $C \in \mathbb{R}^{n\times m}$ with each entry indicating the Manhattan distance between two entities. The global EA task can be formulated into an optimal transport problem (OTP) to find an optimal global alignment by minimizing the total transport distance:

$$\begin{aligned} &\min \sum_{i=1,j=1}^{n,m} C_{ij}\Psi_{ij}, \\ &\text{s.t.} \sum_j \Psi_{ij} = 1, 1 \le i \le n, \\ &\quad \sum_i \Psi_{ij} = 1, 1 \le j \le m, \end{aligned} \tag{5}$$

where Ψ is the transport matrix, and $\Psi_{ij} \in \{0, 1\}$ for all i and j indicates whether entity e_i in $\mathcal{G}_1$ aligns to e_j in $\mathcal{G}_2$. The constraints guarantee the one-to-one alignment. Considering that $n \neq m$ in most cases and the existence of dangling entities, this OTP is invalid. To address these issues, we add an empty entity into $\mathcal{G}_1$ and $\mathcal{G}_2$ separately. Without loss of generality, we prepend the empty entity as the first entity in both KGs. Since we have no information for empty entities, we define hyperparameters, α and β, to describe the cross KG distance between the empty entity and other entities. Therefore, the OTP is now as follow:

$$\begin{aligned} &\min \sum_{i=1,j=1}^{n+1,m+1} C_{ij}\Psi_{ij}, \\ &\text{s.t.} \sum_j \Psi_{ij} = 1, 2 \le i \le n+1, \\ &\quad \sum_i \Psi_{ij} = 1, 2 \le j \le m+1, \end{aligned} \tag{6}$$

where $C_{1,j} = \alpha, \forall j$ and $C_{i,1} = \beta, \forall i$ denote the first row and the first column of the distance matrix, respectively. $\Psi_{ij} \in \{0, 1\}$, and $\Psi_{i,1} = 1$ indicates that entity e_i is dangling, while $\Psi_{1,j} = 1$ also indicates dangling entity e_j. The other $\Psi_{i,j} = 1$ predicts the entity pair (e_i, e_j).

Datasets		#Ent.	#Rel.	#Trip.	#Pairs	#Dang.
MedED	FR	19,382	431	455,368	6,365	13,017
	EN	18,632	622	841,792		12,267
MedED	ES	19,228	546	594,130	11,153	8,075
	EN	18,632	622	841,792		7,479
DBP15K	ZH	19,388	1,700	70,414	15,000	-
	EN	19,572	1,322	95,142		-
DBP15K	JA	19,814	1,298	77,214	15,000	-
	EN	19,780	1,152	93,484		-
DBP15K	FR	19,661	902	105,998	15,000	-
	EN	19,993	1,207	115,722		-

Table 1: Statistics of MedED and DBP15K.

Our approach now merges the EA and the DED into one OTP. This OTP considers the global alignment information and the interactions among alignments and dangling entity identification. Moreover, considering that similar entities contain more information for both EA and DED, we keep the top K rank similarity entities in the other KG for each entity and drop the remaining entities to reduce the complexity of the OTP. Therefore, we solve the problem with very sparse matrices, C and Ψ. Section 5.3 will show that the method is powerful with acceptable computational complexity after reduction. The last problem is to find the proper α and β for both EA and DED. Since we have the pseudo entity pairs set $\mathcal{P}$ in real-world data, we propose an ingenious way to grid search the quantiles of row minimums and column minimums of C synchronously and then select α^* and β^* that achieve the best EA performance on $\mathcal{P}$. Finally, the entities aligned to the empty entity under given α^* and β^* are dangling entities. The other alignments are the global EA results.

4 Experimental Setup

4.1 Datasets and Evaluation

Sun et al. (2021) construct a dataset providing EA task and DED task, which contains the information of relation triples only so that the quality of local EA is limited and therefore incompatible with global alignment methods. In this work, we construct a dataset with graph structure and literal semantic information providing both EA and DED tasks.

Dataset Construction

The Unified Medical Language System (UMLS) (Lindberg et al., 1993) is a large-scale resource containing over 4 million unique medical concepts and over 87 million relation triples. Concepts in UMLS have several terms in different languages. We extract concepts that contain terms in the selected language as entities to construct new monolingual KG and retain the relations between entities. For the entity names, we select the preferred terms in UMLS. The criterion of entity pairs is whether entities belong to the same concept. Similarly, an entity is dangling if its original concept is not in the other KG. We extracted the KGs of English, French, and Spanish and then constructed the KG pairs of FR-EN (French to English) and ES-EN (Spanish to English). We select 20 thousand entities with the most relation triples in UMLS for the specified language and then drop the entities unrelated to other selected entities. Table 1 shows the statistics of the new dataset, MedED. For both EA and DED, we split 70% of entity pairs and dangling entities as the test set. Even though our method does not rely on the training set, we keep the remaining 30% as the training set for further model comparison and ablation study.

DBP15K

We conduct experiments on the widely used existing EA benchmark, DBP15K (Sun et al., 2017). Three pairs of cross-lingual KGs, ZH-EN (Chinese to English), JA-EN (Japanese to English), and FR-EN (French to English), were built into this dataset. Each KG contains approximately 20 thousand entities, and every KG pair contains 15 thousand entity pairs (Table 1). Following the setting in previous works (Sun et al., 2017; Wu et al., 2019; Zhu et al., 2021a), we keep 70% of entity pairs for testing and 30% for training.

Evaluation

We compute two evaluation metrics following previous works for the EA task, Hits@k and mean reciprocal rank (MRR). Hits@k indicates the percentage of the targets that have been correctly ranked in the top K. MRR is the average of the reciprocal of the rank results. The previous EA works compute Hits@k and MRR in a *relaxed setting* in which only the entities in testing pairs are taken into account, assuming that any source entity has a counterpart in the target KG. In addition to the relaxed evaluation, we also compute Hits@k and MRR in a *practical setting* in which for every testing entity, the list of candidate counterparts consists of all entities in the other KG. Global alignment methods generate one-to-one entity pairs, and we evaluate Hits@1 for these methods.

For the DED task, we compute precision, recall, and F1-score for identifying dangling entities.

4.2 Compared Methods

For the EA task, we compare our approach with previous methods we introduced in Section 2: (1) Init-Emb, the initial embeddings used in UED and main comparison models; (2) the methods based on translational KG embeddings model: MTransE (Chen et al., 2017), JAPE (Sun et al., 2017), and BootEA (Sun et al., 2018); (3) the methods based on graph neural networks: RDGCN (Wu et al., 2019), CEA (Zeng et al., 2020), RNM (Zhu et al., 2021b), RAGA (Zhu et al., 2021a), SelfKG (Liu et al., 2021), EchoEA (Lin et al., 2021).

The proposed method is compatible with supervised training entity pairs, so we provide both unsupervised and supervised versions of our method: (1) the unsupervised method, UED, described in Section 3. (2) the supervised version of UED, which combines the training entity pairs and the pseudo entity pairs for the align loss, denoted as UED*.

4.3 Implementation Details

Following Wu et al. (2019), we translate entity names in MedED to English via Google Translate and then use mean of word vector from GloVe (Pennington et al., 2014) to construct the initial entity embeddings. For entities in DBP15K, we inherit the initial embeddings used in previous works (Wu et al., 2019; Zeng et al., 2021; Zhu et al., 2021a,b; Lin et al., 2021). The threshold for pseudo entity pairs ε is 0.99, and the $k = 3$ in globally guided similarity and loss. The initial value of $w(t)$ is 0.3 and $w(t)$ decreases linearly to 0 at 1/4 of the total training steps. We adopt RAGA (Zhu et al., 2021a) as the embedding-based EA model in Section 3.3 to generate enhanced entity embeddings and use the default setting of hyperparameters in RAGA. For α^* and β^* in the global EA and DED, the default value of K is 100 for our method. We grid search 100 paired quantiles of the row minimums and column minimums of C with $K = 10$. Then, α^* and β^* are used in the other values of K.

5 Results

5.1 Entity Alignment Results

Table 2 shows the results of EA on DBP15K and MedED. Following the previous work, we adopt the relaxed evaluation setting. The results with practical evaluation setting are listed in Appendix A.1.

In general, for both local and global alignment in DBP15K, the UED achieves comparable results with the previous state-of-the-art baselines. More specifically, for local alignment, the UED achieves the same level behavior as the supervised embedding-based EA method, the RAGA, of which we adopt its graph embedding models. For global alignment, the OTP brings UED a significant improvement, and the UED outperforms all competing methods except the new supervised state-of-the-art method, EchoEA. The Hits@1 of UED for ZH-EN, JA-EN, and FR-EN achieves 0.877, 0.915, and 0.975 in DBP15K, respectively. In addition, UED* outperforms all methods and achieves 0.915 and 0.941 Hits@1 for ZH-EN and JA-EN in DBP15K and 0.974 and 0.979 for FR-EN and ES-EN in MedED.

5.2 Entity Alignment and Dangling Entity Detection Results

Table 3 shows the results of EA and DED on MedED. Note that global alignment with DED should consider all entities. We select the practical setting in the EA evaluation.

As shown in Table 3, for the EA task, by maximizing the performance of EA on pseudo entity pairs, UED achieves better results compared to the supervised RAGA and the variants of our method with DAA. In addition, the UED ($K = 100$) achieves 0.805 and 0.877 Hits@1 for FR-EN and ES-EN separately. The supervised UED* gains a further improvement of 0.021 and 0.012 Hits@1 for FR-EN and ES-EN separately. For the DED task, the proposed method focuses more on the precision in recognizing dangling entities. The results of UED and UED* are also much better than the Distance. The Distance denotes the baseline by searching the best threshold on the dangling training set for identifying dangling entities according to the smallest distance to entities in another KG. These results imply that UED successfully uses unsupervised EA to assist DED while DED with high precision reduces the scope of EA and enhances the performance of EA. Furthermore, the results with different K show that we don't need a vary large value of K, and there is a tradeoff between improving EA results and DED results: the larger K achieves the better Hits@1 in the EA task and precision in the DED task, while the smaller K

	DBP15K									MedED					
	ZH-EN			JA-EN			FR-EN			FR-EN			ES-EN		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
Local															
Init-Emb	.575	.689	.615	.650	.754	.688	.818	.888	.843	.716	.845	.764	.685	.826	.737
MTransE	.308	.614	.364	.279	.575	.349	.244	.556	.335	-	-	-	-	-	-
JAPE	.731	.904	-	.828	.947	-	-	-	-	-	-	-	-	-	-
BootEA	.629	.848	.703	.622	.854	.701	.653	.874	.731	-	-	-	-	-	-
RDGCN	.708	.846	-	.767	.895	-	.886	.957	-	-	-	-	-	-	-
RNM	.840	.919	**.870**	.872	.944	.899	.938	.981	.954		-	-	-	-	-
RAGA	.798	.930	.847	.831	.950	.875	.914	.983	.940	.896	**.981**	.930	**.914**	.986	**.943**
SelfKG	.829	.919	-	.890	.953	-	.959	**.992**	-	-	**-**	-	**-**	-	**-**
EchoEA	.823	.939	.865	.861	.957	.897	.939	.989	.958	-	**-**	-	**-**	-	**-**
UED	.779	.907	.826	.820	.933	.862	.921	.979	.943	.895	.975	.926	.893	.978	.925
UED*	.826	**.943**	**.870**	.863	.960	**.900**	.938	.987	**.957**	.901	**.981**	**.932**	.913	**.987**	.942
Global															
GM-EHD-JEA	.736			.792			.924			-			-		
CEA	.787			.863			.972			-			-		
RAGA	.873			.909			.966			.962			.970		
EchoEA	.891			.932			**.989**			-			-		
UED	.877			.915			.975			.970			.976		
UED*	**.915**			**.941**			.984			**.974**			**.979**		

Table 2: EA results on DBP15K and MedED datasets (relaxed setting). H@1 and H@10 denotes the Hits@1 and Hits@10. The underlined models use the same initial entity embeddings. The results of the compared method in DBP15K are from their original papers. We apply the RAGA in MedED for comparison. The CEA, RAGA and EchoEA use the DAA for global alignment.

	FR-EN				ES-EN			
	EA	DED			EA	DED		
	H@1	P	R	F	H@1	P	R	F
RAGA	.787	-	-	-	.827	-	-	-
UED(DAA)	.774	-	-	-	.870	-	-	-
Distance	-	.781	.734	.757	-	.786	**.861**	.822
UED								
K=1	.798	.961	**.794**	**.869**	.860	.904	.842	**.872**
K=10	.803	.963	.753	.845	.874	.935	.684	.790
K=100	.805	.964	.748	.842	.877	.933	.646	.764
UED*	**.826**	**.976**	.654	.783	**.901**	**.941**	.694	.799

Table 3: EA and DED results on MedED (practical setting). H@1, P, R, and F denotes Hits@1, precision, recall, and F-score. $K = 1, 10, 100$ refers to the proposed global alignment method that keeps the top $K(= 1, 10, 100)$ rank similarity entities for each entity. The UED(DAA) and RAGA use the DAA for global alignment.

achieves the better F1-score in the DED task.

5.3 Empirical Runtime Analysis

The time complexity of the proposed global method is acceptable. The solving process of the OTP could be finished in less than 7, 60, and 5,00 seconds for $K = 1, 10, 100$ in MedED. Without the simplification, the running time will be more than 120,000 seconds. Considering the time consuming and the similar performance of $K = 10$ and $K = 100$ (Table 3), much larger value of K may not bring significant improvement and $K = 100$ is enough for the proposed method.

6 Ablation Study

To quantify the role of our designs, we provide the variants by removing the weight decreasing mechanism of the globally guided loss $\mathcal{L}_g$ and the $\mathcal{L}_g$ from UED (Table 4). In addition, we attempt to replace the proposed OTP with DAA (Table 4). For local alignment, the UED without $\mathcal{L}_g$ is the same as RAGA except for the training entity pairs. Table 5 provides other necessary results and variants in practical setting. There are five major observations:

1. The performance of our method with pseudo entity pairs is similar to those with true entity pairs. For example, in Table 4, for local alignment results of FR-EN in DBP15K, the UED without $\mathcal{L}_g$ uses 10,689 pseudo entity pairs and gains 0.913 Hits@1, while the RAGA uses 4500 true entity pairs and gains 0.914 Hits@1. Although the proportion of how many pseudo entity pairs can play an equal role as true entity pairs changes, depending on the quality of the initial entity embedding and the KGs (Figure 2), it is valid to obtain pseudo entity pairs when true entity pairs are unavailable.

	DBP15K			MedED	
	ZH	JA	FR	FR	ES
Local					
RAGA	.798	.831	.914	.896	.914
UED	.779	.820	.929	.895	.893
w/o $\mathcal{L}_g$	.759	.794	.913	.891	.896
Global					
UED	.877	.915	.975	.970	.976
w/o dec.	.873	.910	.973	.969	.973
w/o $\mathcal{L}_g$	.875	.910	.973	.971	.975
w/o OTP	.779	.820	.921	.895	.893
UED(DAA)	.847	.891	.962	.955	.956

Table 4: Hits@1 results of method variants (relaxed setting) in the EA task. The dec. is the weight decreasing mechanism of the globally guided loss, $\mathcal{L}_g$. ZH, JA, FR and ES denotes the KG pairs ZH-EN, JA-EN, FR-EN and ES-EN.

	FR-EN		ES-EN	
	EA	DED	EA	DED
UED	.803	.845	.874	.790
w/o empty	.555	-	.652	-
w. gold α, β	.809	.803	.874	.790
UED(CODER)	.884	.863	.933	.865

Table 5: Results of method variants (practical setting) in MedED. We report Hits@1 and F-score for EA and DED. The w/o empty denotes the OTP without the empty entities. The w. gold α, β denote that the α and β in the OTP are selected by the dangling training set. UED(CODER) refers to the method that we replace the Glove with a medical language model in UED.

2. The proposed global alignment method is stable and effective, causing significant improvements (0.046∼0.098 Hits@1) compared with the UED for local alignment Table 4).

3. The globally guided similarity and loss and the weight decreasing mechanism are usually helpful (Table 4).

4. Introducing the empty entity is necessary. The global method without empty entities harms the EA result and cannot be applied to the DAD task (Table 5).

5. The proposed method for searching proper α^* and β^* produces successful results. The results with α^* and β^* achieve the same level of performance for EA and DED compared to the gold selection for α and β based on the EA training entity pairs.

Besides, we attempt to replace the GloVe in MedED with a pretrained medical language model (LM), the English version of CODER (Yuan et al., 2022), and show that a proper domain-specific LM trained on a large KG may achieve better results (Table 5).

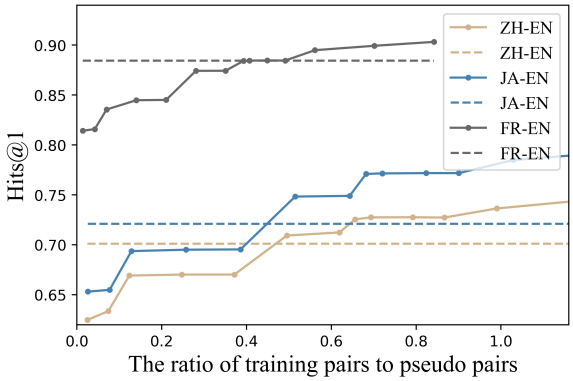

Figure 2: The Hits@1 in DBP15K (practical setting) for the UED without $\mathcal{L}_g$ and OTP. The solid line and dashed line denotes the method trained with the training entity pairs and pseudo entity pairs, respectively.

7 Conclusion

This paper proposes a novel unified unsupervised method for both EA and DED, which better fits the realistic scenario for integrating KGs. UED contains four modules: pseudo entity pair generation, information aggregation, globally guided similarity and loss, and a modified OTP for global EA and DED. The first three modules mine the information in KGs to get rid of supervised entity pairs, while the last module integrates EA and DED into a unified framework to identify dangling entities without supervision and provide better EA results. We also construct a new dataset for the EA and DED tasks and perform experiments to demonstrate the effectiveness of UED.

Acknowledgements

We would like to express our gratitude to the reviewers for their helpful comments and suggestions. We thank Zheng Yuan, Hongyi Yuan, Pengyu Cheng, and Huaiyuan Ying for their help.

References

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Yixin Cao, Xiang Wang, Xiangnan He, Zikun Hu, and Tat-Seng Chua. 2019. Unifying knowledge graph

learning and recommendation: Towards a better understanding of user preferences. In *The world wide web conference*, pages 151–161.

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

Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys (CSUR)*, 55(2):1–36.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Xueyuan Lin, Wenyu Song, Haoran Luo, et al. 2021. Echoea: Echo information between entities and relations for entity alignment. *arXiv preprint arXiv:2107.03054*.

Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The unified medical language system. *Yearbook of Medical Informatics*, 2(01):41–51.

Xiao Liu, Haoyun Hong, Xinghao Wang, Zeyi Chen, Evgeny Kharlamov, Yuxiao Dong, and Jie Tang. 2021. A self-supervised method for entity alignment. *arXiv preprint arXiv:2106.09395*.

Xin Mao, Wenting Wang, Huimin Xu, Yuanbin Wu, and Man Lan. 2020. Relational reflection entity alignment. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1095–1104.

Tam Thanh Nguyen, Thanh Trung Huynh, Hongzhi Yin, Vinh Van Tong, Darnbi Sakong, Bolong Zheng, and Quoc Viet Hung Nguyen. 2020. Entity alignment for knowledge graphs with multi-order convolutional networks. *IEEE Transactions on Knowledge and Data Engineering*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Denis Savenkov and Eugene Agichtein. 2016. Crqa: Crowd-powered real-time automatic question answering system. In *Fourth AAAI conference on human computation and crowdsourcing*.

Zequn Sun, Muhao Chen, and Wei Hu. 2021. Knowing the no-match: Entity alignment with dangling cases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3582–3593, Online. Association for Computational Linguistics.

Zequn Sun, Wei Hu, and Chengkai Li. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *International Semantic Web Conference*, pages 628–644. Springer.

Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. 2018. Bootstrapping entity alignment with knowledge graph embedding. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4396–4402.

Bayu Distiawan Trisedya, Jianzhong Qi, and Rui Zhang. 2019. Entity alignment between knowledge graphs using attribute embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 297–304.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 349–357.

Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, Rui Yan, and Dongyang Zhao. 2019. Relation-aware entity alignment for heterogeneous knowledge graphs. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence.

Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web*, pages 1271–1279.

Kun Xu, Linfeng Song, Yansong Feng, Yan Song, and Dong Yu. 2020. Coordinated reasoning for cross-lingual knowledge graph alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9354–9361.

Kun Xu, Liwei Wang, Mo Yu, Yansong Feng, Yan Song, Zhiguo Wang, and Dong Yu. 2019. Cross-lingual knowledge graph alignment via graph matching neural network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3156–3161, Florence, Italy. Association for Computational Linguistics.

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2017. Improved neural relation detection for knowledge base question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 571–581, Vancouver, Canada. Association for Computational Linguistics.

Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2022. Coder: Knowledge-infused cross-lingual medical term embedding for term normalization. *Journal of biomedical informatics*, page 103983.

Weixin Zeng, Xiang Zhao, Jiuyang Tang, Xinyi Li, Minnan Luo, and Qinghua Zheng. 2021. Towards entity alignment in the open world: an unsupervised approach. In *International Conference on Database Systems for Advanced Applications*, pages 272–289. Springer.

Weixin Zeng, Xiang Zhao, Jiuyang Tang, and Xuemin Lin. 2020. Collective entity alignment via adaptive features. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1870–1873. IEEE.

Xiang Zhao, Weixin Zeng, Jiuyang Tang, Wei Wang, and Fabian Suchanek. 2020. An experimental study of state-of-the-art entity alignment approaches. *IEEE Transactions on Knowledge & Data Engineering*, (01):1–1.

Renbo Zhu, Meng Ma, and Ping Wang. 2021a. Raga: Relation-aware graph attention networks for global entity alignment. In *PAKDD (1)*, pages 501–513. Springer.

Yao Zhu, Hongzhi Liu, Zhonghai Wu, and Yingpeng Du. 2021b. Relation-aware neighborhood matching model for entity alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4749–4756.

A Appendix

A.1 Practiacl Evaluation Results

For completeness, this appendix reports the EA results on DBP15K in practiacl evaluation setting (Table 6).We compared our methods with the RAGA, since we adopt the part of graph embedding in RAGA in our framework.

	ZH-EN			JA-EN			FR-EN		
	@1	@10	MRR	@1	@10	MRR	@1	@10	MRR
local									
Init-Emb	.570	.686	.611	.633	.753	.676	.807	.890	.835
RAGA	.725	.903	.790	.773	.931	.829	.884	.972	.917
UED	.751	.892	.802	.793	.918	.839	.911	.974	.934
global									
RAGA	.834			.742			.929		
UED(DAA)	.799			.769			.935		
UED	**.847**			**.890**			**.966**		

Table 6: EA results on DBP15K (practical setting). @1 and @10 denotes the Hits@1 and Hits@10. The UED(DAA) refer to the variant of UED by replacing the OTP with DAA.