DAEA: Enhancing Entity Alignment in Real-World Knowledge Graphs Through Multi-Source Domain Adaptation

Linyan Yang[1,2], Shiqiao Zhou[2]*, Jingwei Cheng[1,3], Fu Zhang[1,3], Jizheng Wan[2], Shou Wang[2], Mark Lee[2]

[1]School of Computer Science and Engineering, Northeastern University, China
[2]School of Computer Science, University of Birmingham, UK
[3]Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education, Northeastern University, China

yanglinyanly@163.com, sxz363@student.bham.ac.uk, {chengjingwei, zhangfu}@mail.neu.edu.cn, {j.wan.1, s.wang.2, m.g.lee}@bham.ac.uk

Abstract

Entity Alignment (EA) is a critical task in Knowledge Graph (KG) integration, aimed at identifying and matching equivalent entities that represent the same real-world objects. While EA methods based on knowledge representation learning have shown strong performance on synthetic benchmark datasets such as DBP15K, their effectiveness significantly decline in real-world scenarios which often involve data that is highly heterogeneous, incomplete, and domain-specific, as seen in datasets like DOREMUS and AGROLD. Addressing this challenge, we propose DAEA, a novel EA approach with **D**omain **A**daptation that leverages the data characteristics of synthetic benchmarks for improved performance in real-world datasets. DAEA introduces a multi-source KGs selection mechanism and a specialized domain adaptive entity alignment loss function to bridge the gap between real-world data and optimal benchmark data, mitigating the challenges posed by aligning entities across highly heterogeneous KGs. Experimental results demonstrate that DAEA outperforms state-of-the-art models on real-world datasets, achieving a 29.94% improvement in Hits@1 on DOREMUS and a 5.64% improvement on AGROLD. Code is available at https://github.com/yangxiaoxiaoly/DAEA.

1 Introduction

Knowledge Graphs (KGs) have recently been developed and utilized across various domains. However, since most KGs are created independently by different organizations and individuals, they often exhibit significant heterogeneity. Knowledge fusion seeks to address this by aligning and merging heterogeneous and redundant information within KGs to achieve a globally unified representation of knowledge (Dong et al., 2014). Entity Alignment (EA) plays a crucial role in this fusion process, with its primary objective being to identify equivalent entities across different KGs.

In recent years, methods based on knowledge representation learning have become increasingly popular for tackling the entity alignment challenge. These methods work by projecting entities into a low-dimensional vector space, where the similarity between entities is determined based on their embeddings. MTransE (Chen et al., 2017), BootEA (Sun et al., 2018), JAPE (Sun et al., 2017), and TransEdge (Sun et al., 2019) utilize TransE (Bordes et al., 2013) to learn entity and relation embeddings. GNN-based EA methods (Wang et al., 2018; Xu et al., 2019; Wu et al., 2019a) generate entity embeddings by aggregation information from their neighbourhoods via GNNs (Kipf and Welling, 2017). These methods are based on the premise that similar neighborhood structures exist in different KGs, implying isomorphism, which may not hold true due to the heterogeneity of KGs (Sun et al., 2020). To address this, some approaches have applied an attention mechanism to weigh relations between entities differently (Mao et al., 2020; Wu et al., 2019a) or have selectively ignored neighbors that are detrimental to alignment (Wu et al., 2020; Cao et al., 2019; Li et al., 2019). Additionally, attributes of triples have been recognized as vital for alignment. Several strategies enhance alignment by embedding attributes such as names, types, or values alongside structural embeddings (Sun et al., 2017; Wang et al., 2018; Chen et al., 2020; Zhang et al., 2019; Trisedya et al., 2019; Wang et al., 2020; Tang et al., 2020; Zhong et al., 2022).

In real-world datasets, issues such as high heterogeneity, sparsity, and incompleteness are prevalent (Lisena et al., 2018; Venkatesan et al., 2018). Not only many corresponding entities have completely different neighbors (as illustrated in Figure 1), but numerous entities also lack attribute information.

*Corresponding Author
[1]https://www.wikidata.org/wiki/Wikidata:Main_Page
[2]https://www.dbpedia.org/

Proceedings of the 31st International Conference on Computational Linguistics, pages 5890–5901
January 19–24, 2025.

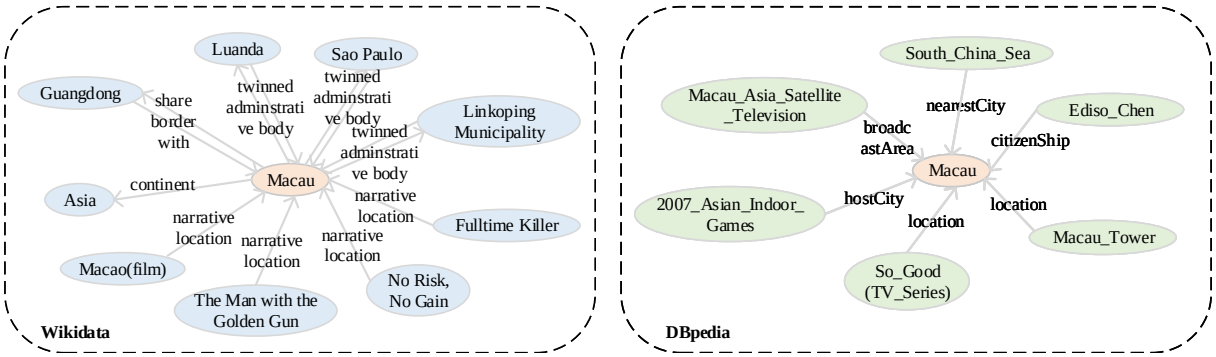

Figure 1: An example of entity alignment in real-world KGs (Wikidata [1]and DBpedia[2]). The yellow backgrounds represent the same entities in two KGs. The surrounding blue and green backgrounds represent their neighbor entities, while the solid lines with arrows represent the relations between entities.

Therefore, relying solely on the information inherent within KGs is insufficient for effectively learning and performing EA. This limitation significantly leads to the degradation in performance of these models when applied to real-world datasets.

To enhance EA performance on real-world datasets, we propose the **D**omain **A**daptive **E**ntity **A**lignment (DAEA) method. This innovative approach aims to enhance the model's adaptability and accuracy in diverse real-world environments by leveraging rich knowledge from source datasets. We first propose a multi-source KGs selection mechanism that strategically selects relevant dataset from multiple source KGs. If source data is selected for transfer learning without careful consideration, it may cause negative transfer. Therefore, the mechanism selects the source KGs that are most similar to the target KGs for domain adaptation, taking into account both semantic and structural information. By incorporating insights gained from synthetic benchmarks, the mechanism strengthens the model's ability to align entities more accurately, even in the face of complex and diverse KGs.

Additionally, we design a domain adaptive entity alignment loss function that plays a crucial role during the training phase of the model by reducing the distance between corresponding entities, thereby aligning them more closely. Simultaneously, it also works to minimize the distributional disparities between benchmark data, which is often idealized or standardized, and real-world data, which contains more variability and noise. In summary, the main contributions of this paper are as follows:

- We propose a multi-source KGs selection mechanism that fully leverages the valuable information available in benchmark datasets to enhance entity alignment in real-world datasets.
- We design a domain adaptive entity alignment loss function with a dual focus on both entity alignment and domain adaptation, which helps in achieving a more holistic improvement in model performance.
- To the best of our knowledge, this is the first instance of applying domain adaptation techniques from transfer learning specifically to the task of EA. Extensive experiments demonstrate that our DAEA method outperforms SOTA models on real-world datasets.

2 Related Work

2.1 Entity alignment

Currently, the majority of Entity Alignment (EA) methods are grounded in knowledge representation learning, and can be primarily categorized into either translation based methods or based on GNNs/GCNs. Translation based methods, such as MTransE (Chen et al., 2017), JAPE (Sun et al., 2017), KECG (Li et al., 2019), BootEA (Sun et al., 2018), Multi-mapping Relations (Shi and Xiao, 2019), TransEdge (Sun et al., 2019), JarKA(Chen et al., 2020), and CTEA(Yan et al., 2020), principally constrain the entity embeddings into a fixed distribution by translation-based knowledge graphs embedding methods. Based on the observation that entities sharing similar neighboring structures tend to be aligned, EA approaches based on GCNs distribute and consolidate entity information across graphs. GCN-Align (Wang et al., 2018) is the first to use GCN to jointly embed the entity structure and entity attributes. Building upon this foundation, many approaches have enhanced GCNs to address issues such as noise propagation (HGCN

(Wu et al., 2019b)), heterogeneity (MuGNN (Cao et al., 2019), Alinet (Sun et al., 2020), NMN (Wu et al., 2020), MRAEA (Mao et al., 2020)), and better utilization of relationship and attribute information (RDGCN (Wu et al., 2019a), RAGA (Zhu et al., 2021a), RNM (Zhu et al., 2021b), EPEA (Wang et al., 2020)). The ExEA (Tian et al., 2024) framework is designed to generate high-quality explanations for a given embedding-based EA model while also improving EA results through repair. CAECGAT (Xie et al., 2021) jointly learns cross-KG embeddings by propagating information across different KGs, and DuGa-DIT (Xie et al., 2020) bridges the semantic gap between KGs by leveraging both neighborhood features and cross-KG alignment information. And some temporal entity alignment methods, like TS-align (Zhang et al., 2024), Simple-HHEA (Jiang et al., 2024). Simple-HHEA highlights the challenge of entity alignment in heterogeneous knowledge graphs and introduces a new time-aware heterogeneous knowledge graph entity alignment dataset.

With the rise of pre-trained language models like BERT (Kenton and Toutanova, 2019), HMAN+BERT (Yang et al., 2019), SDEA (Zhong et al., 2022), and BERT-INT (Tang et al., 2020) treat entity alignment as a downstream task for fine-tuning BERT.

Due to the high heterogeneity and limited available information in real-world datasets, existing entity alignment methods, despite showing superior performance on benchmarks, experience significant performance degradation when applied to real-world datasets. Consequently, we propose DAEA approach, which incorporates domain adaptation techniques to enhance entity alignment performance in real-world datasets.

2.2 Domain Adaptation

Domain adaptation (DA) is a key area within transfer learning (Pan and Yang, 2009), aiming to adapt models from a source domain to a target domain with differing distributions. Techniques in DA focus on extracting domain-invariant representations by utilizing distance metrics like Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) and Correlation Alignment (CORAL) (Sun and Saenko, 2016), as well as employing adversarial methods such as Domain-adversarial Neural Network (DANN) (Ganin et al., 2016) and Multi-adversarial Domain Adaptation (MADA) (Pei et al., 2018) to align these distributions. For more complex scenarios involving multiple sources, multi-source domain adaptation (MDA) is necessary. In entity linkage topic, AdaMEL (Jin et al., 2021) leverages attribute information to adapt labeled data from different source datasets to target dataset. Research in multi-source graph domain adaptation (GDA) includes models like NESTL (Fu et al., 2020), which trains individual models for each source based on topological similarity, and MSDS (He et al., 2023), which selects the most transferable sources using mixed discrepancy metrics. Additionally, Meta-GDN (Ding et al., 2021) facilitates few-shot network anomaly detection by transferring meta-knowledge from multiple networks.

Although there has been much research on GDA, there has been no research on domain adaptation for entity alignment where the datasets not only contain graph pairs, but also have heterogeneous structures, presenting more challenges.

3 Task Definition

Definition 1 (Knowledge Graph) A knowledge graph (KG) is denoted as $G = (E, R, A, V, T_r, T_a)$, where $E = \{e_1, e_2, ...e_m\}$, $R = \{r_1, r_2, ...r_n\}$, $A = \{a_1, a_2, ...a_p\}$, and $V = \{v_1, v_2, ..., v_q\}$ represent entity set, relation set, attribute set, and value set, respectively, and m, n, p, q are the number of entities, relations, attributes, and attribute values, respectively. $T_r \subseteq E \times R \times E$ is the relation triple set, and $T_a \subseteq E \times A \times V$ is the attribute triple set. Relational triples can also be represented as (h, r, t), where h is called the head entity and t is called the tail entity.

Definition 2 (Entity Alignment in KGs) Given two KGs $G^1 = (E^1, R^1, A^1, V^1, T_r^1, T_a^1)$, and $G^2 = (E^2, R^2, A^2, V^2, T_r^2, T_a^2)$, the aligned entity pairs (training set) is denoted as $S = \{(e_i^1, e_j^2) | e_i^1 \in E^1, e_j^2 \in E^2, e_i^1 \equiv e_j^2\}$, where $\equiv$ stands for equivalence, i.e., the entity e_i^1 and entity e_j^2 refer to the same thing in the real world. The goal of the EA task is to find the remaining equivalent entity pairs of these two KGs.

Definition 3 (Source and Target KGs) The source KGs G_{s_l} refers to a graph pairs $\{(G_{s_l}^1, G_{s_l}^2)\}$. Here l means source dataset number. The superscript of graph pairs means the order of graph. There are multiple source KGs: $GS = \{G_{s_1} = (G_{s_1}^1, G_{s_1}^2), ..., G_{s_u} = (G_{s_u}^1, G_{s_u}^2)\}$, and target KGs $GT = (G_t^1, G_t^2)$, u is the number of source KGs, where

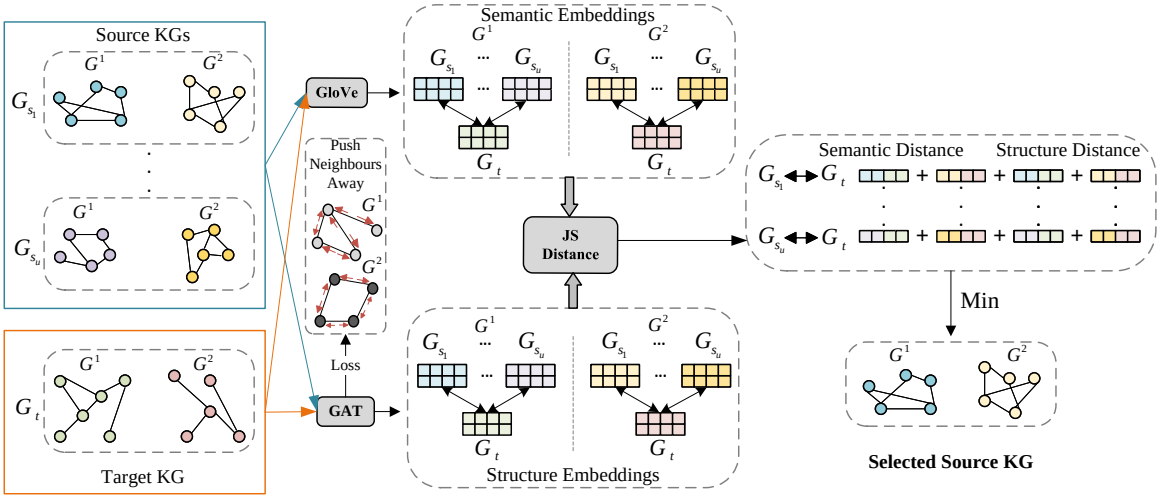

Figure 2: Multi-source selection

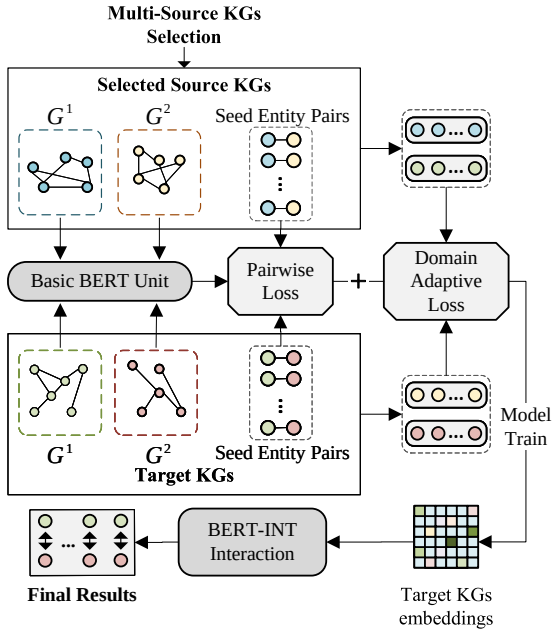

Figure 3: Domain Adaptation

$G^1_{s_1} = (E^1_{s_1}, R^1_{s_1}, A^1_{s_1}, V^1_{s_1}, T^1_{rs_1}, T^1_{as_1})$ and $G^1_t = (E^1_t, R^1_t, A^1_t, V^1_t, T^1_{rt}, T^1_{at})$. The aligned entity pairs (training set) are denoted as $S_{s_l} = \{(e^1_{is_l}, e^2_{js_l}) | e^1_{is_l} \in E^1_{s_l}, e^2_{js_l} \in E^2_{s_l}, e^1_{is_l} \equiv e^2_{js_l}\}$ and $S_t = \{(e^1_{it}, e^2_{jt}) | e^1_{it} \in E^1_t, e^2_{jt} \in E^2_t, e^1_{it} \equiv e^2_{it}\}$, where $\equiv$ stands for equivalence.

4 Methodology

The DAEA model primarily comprises of two components: Multi-Source KGs Selection (Figure 2) and Domain Adaptation (Figure 3).

Multi-Source KGs Selection is employed to identify which source KGs from the benchmark are most suitable to do transfer learning to the target KGs. In Figure 2, the optimal KGs for transfer learning are selected by calculating the semantic and structural distances between various KGs. Semantic and structural information are captured using GloVe (Pennington et al., 2014) embeddings and Graph Attention Networks (GATs) (Velickovic et al., 2017) respectively, with the latter employing an unsupervised contrastive learning loss. A more detailed discussion will be provided in Section 4.1.

Figure 3 details the process of domain adaptation. Initially, data input is expanded on the basis of BERT-INT architecture to include both source and target KGs. During the training phase, the model not only employs pairwise margin loss to approximate the corresponding entities in the source and target KGs but also computes domain adaptive loss between the training sets of the source and target KGs. More details will be discussed in Section 4.2.

4.1 Multi-source KGs selection

To more comprehensively assess the similarity between KGs, we consider both semantic and structural information. Let $\mathcal{D}_{G_sG_t}$ represent the distance between the source KGs (GS) and the target KGs (GT). Specifically, we define $\mathcal{D}_{G_sG_t}$ as follows:

$$\mathcal{D}_{G_sG_t} = \{\mathcal{D}_{G_{s_1}G_t}, \ldots, \mathcal{D}_{G_{s_u}G_t}\} \quad (1)$$

The smaller the distance between source and target KGs, the higher their similarity. Therefore, we select the source KGs with the smallest distance as the optimal. $\mathcal{D}_{G_{s_i}G_t}$ for each individual component $i = 1, \ldots, u$ is given by:

$$\mathcal{D}_{G_{s_i}G_t} = dSE_{G_{s_i}G_t} + dST_{G_{s_i}G_t} \quad (2)$$

Here, $dSE_{G_{s_i}G_t}$ and $dST_{G_{s_i}G_t}$ represent the semantic and structural distances, respectively, and are computed as:

$$dSE_{G_{s_i}G_t} = dSE_{G^1_{s_i}G^1_t} + dSE_{G^2_{s_i}G^2_t} \quad (3)$$

$$dST_{G_{s_i}G_t} = dST_{G^1_{s_i}G^1_t} + dST_{G^2_{s_i}G^2_t} \quad (4)$$

4.1.1 Semantic Distance

We employ the widely-used word embedding tool, GloVe, to obtain the embedding representations of entity names within the KGs. The semantic embedding representations of the source and target KGs are denoted by $SE_{G^1_{s_i}}$, $SE_{G^2_{s_i}}$, $SE_{G^1_t}$, $SE_{G^2_t}$, respectively. We utilize the Jensen-Shannon (JS) distance (Fuglede and Topsoe, 2004), a widely adopted metric, to assess similarities across KGs. $dSE_{G^1_{s_i}G^1_t}$ is computed as:

$$dSE_{G^1_{s_i}G^1_t} = \sqrt{JS(SE_{G^1_{s_i}}, SE_{G^1_t})} \quad (5)$$

where $JS(SE_{G^1_{s_i}}, SE_{G^1_t})$ can be computed as :

$$\begin{aligned} &JS(SE_{G^1_{s_i}}, SE_{G^1_t}) = \\ &\frac{1}{2}D(SE_{G^1_{s_i}} \parallel \mathcal{M}) + \frac{1}{2}D(SE_{G^1_t} \parallel \mathcal{M}) \\ &\mathcal{M} = \frac{1}{2}(SE_{G^1_{s_i}} + SE_{G^1_t}) \end{aligned} \quad (6)$$

Here, $D(P \parallel Q)$ for P and Q, can be computed as:

$$D(P \parallel Q) = \sum_i P_i log(\frac{P_i}{Q_i}) \quad (7)$$

The calculation method for $dSE_{G^2_{s_i}G^2_t}$ follows the same approach.

4.1.2 Structural Distance

A two-layer GAT is employed to extract the structural information from KGs. Specifically, with a standard GAT layer, the hidden state h_i for entity e_i at each layer is performed as follows.

$$h_i = ReLU(\sum_{j \in \mathcal{N}_i} \alpha_{ij} W h_j) \quad (8)$$

where $\mathcal{N}_i$ denotes the set of neighbors of e_i, h_j denotes the embedding of entity e_j obtained by this layer, W is a trainable weight matrix, α_{ij} are the attention coefficients computed as:

$$\alpha_{ij} = \frac{exp(\Gamma(a^T[We_i \oplus We_j]))}{\sum_{k \in \mathcal{N}_i} exp(\Gamma(a^T[We_i \oplus We_k]))} \quad (9)$$

where Γ is the LeakyReLU nonlinear activation function, a is a trainable parameter, $\oplus$ denotes the concatenation operation.

In order to better accommodate the entity alignment task, we utilize an unsupervised contrastive learning loss during the training process when applying GAT to individual graph data. For each entity, the goal is to maximize the distance between it and its neighbouring entities.

$$\mathcal{L}_c = \frac{1}{|\mathcal{N}_i|} \sum_{e_j \in \mathcal{N}_i} max(0, M - Eu(e_i, e_j)) \quad (10)$$

where $|\mathcal{N}_i|$ is the number of $\mathcal{N}_i$, and M is the margin, Eu is the Euclidean distance.

Structural embeddings, denoted as $GAT_{G^1_{s_i}}$, $GAT_{G^1_{s_i}}$, can be obtained through the trained GAT. Subsequently, the structural distance $dST_{G^1_{s_i}G^1_t}$ can be calculated as described in Equation (5).

4.2 Domain Adaptation

In the domain adaptation stage, DAEA follows BERT-INT and treats entity alignment as the downstream task to fine-tune a pre-trained BERT model. Initially, we expand the input data into source KGs and target KGs. Subsequently, we compute the pairwise losses for both the source KGs and the target KGs, as well as the domain adaptive loss between the source KGs and the target KGs. The sum of these three losses constitutes the total loss of the entire model. It can be denoted as:

$$Loss = \mathcal{L}_s + \mathcal{L}_t + \mathcal{L}_{DA} \quad (11)$$

4.2.1 Pairwise Loss

For each entity pairs (e^1_i, e^2_j) in training set S, for entity e^1_i, we treat e^2_j as a positive example, and a negative example e^{2-}_j randomly sampled from E^2. Let $\mathcal{L}_s$ and $\mathcal{L}_t$ respectively represent the pairwise loss for the source KGs and the target KGs. $\mathcal{L}_s$ can be computed as follows.

$$\begin{aligned} \mathcal{L}_s = &\sum_{(e^1_i, e^2_j, e^{2-}_j) \in S} \\ &max\{0, l_1(e^1_i, e^2_j) - l_1(e^1_i, e^{2-}_j) + M\} \end{aligned} \quad (12)$$

where M is the margin, $l_1(e^1_i, e^2_j)$ is the L1 distance between e^1_i and e^2_j. The calculation of $\mathcal{L}_t$ follows the same methodology as above.

4.2.2 Domain Adaptive Loss

We compute the distribution distance between the training sets of the source KGs and the target KGs to serve as the domain adaptive loss.

Given a source training set $S_s = \{(e_{is}^1, e_{is}^2, e_{is}^{2-})|e_{is}^1 \in E_s^1, e_{is}^2 \in E_s^2, e_{is}^{2-} \in E_s^2, \}$ and a target training set $S_t = \{(e_{jt}^1, e_{jt}^2, e_{jt}^{2-})|e_{jt}^1 \in E_t^1, e_{jt}^2 \in E_t^2, e_{jt}^{2-} \in E_t^2, \}$. When computing the domain adaptive loss, we minimize the distance between positive examples from the source and target KGs, as well as the distance between negative examples from the source and target KGs. Let $\mathcal{DA}_P$ and Let $\mathcal{DA}_N$ denote the domain adaptive loss of one positive and negative pairs, respectively. They are denoted as:

$$\mathcal{DA}_P = \sum_{i=1}^{|S_s|} \sum_{j=1}^{|S_t|} (d(e_{is}^1, e_{jt}^1) + d(e_{is}^2, e_{jt}^2)) \quad (13)$$

$$\mathcal{DA}_N = \sum_{i=1}^{|S_s|} \sum_{j=1}^{|S_t|} d(e_{is}^{2-}, e_{jt}^{2-}) \quad (14)$$

where $|\cdot|$ represents the size of a set.

To effectively measure the distance between source and target distributions, we employ MMD (Gretton et al., 2012), which is one of the most widely used metrics in domain adaptation. $d(e_{is}^1, e_{jt}^1)$ is computed as:

$$d(e_{is}^1, e_{jt}^1) = \mathbb{E}[k(e_{is}^1, e_{is}^{1'})] + \mathbb{E}[k(e_{jt}^1, e_{jt}^{1'})] - 2\mathbb{E}[k(e_{is}^1, e_{jt}^1)] \quad (15)$$

where k refers to kernel function, which is Gaussian kernel (Elen et al., 2022) in our case. $e_{is}^{1'}$ and $e_{jt}^{1'}$ are samples from source and target. $\mathbb{E}$ is the expected value. $d(e_{is}^2, e_{jt}^2)$ and $d(e_{is}^{2-}, e_{jt}^{2-})$ are computed with a similar way. Eventually, the domain adaptive loss is denoted as:

$$\mathcal{L}_{DA} = \mathcal{DA}_P + \mathcal{DA}_N \quad (16)$$

5 Experiment

5.1 Experiment Settings

5.1.1 Datasets

The widely used benchmark, DBP15K, is considered as an ideal synthetic dataset for entity alignment, comprising three multilingual sub-datasets: ZH-EN, JA-EN, and FR-EN. In this study, we adopt DBP15K as the source KGs, while utilizing two real-world datasets, DOREMUS (Lisena et al., 2018) and AgroLD (Venkatesan et al., 2018), as the target KGs, which are introduced by (Raoufi et al.)[3]. DOREMUS is a multilingual dataset focused on classical music, and AGROLD is a large dataset for plant science. From Table 1, it is evident that nearly 80% of entities can find their corresponding counterparts in DBP15K, whereas only about 10% of entities have aligned counterparts in real-world data.

Datasets		Entities	Rel.	Rel.Triples	Attr.	Attr.Triples	Pairs
DBP15K							
ZH-EN	ZH	19388	1701	70414	7780	379684	15000
	EN	19572	1323	95142	6933	567755	
JA-EN	JA	19814	1299	77241	5681	354619	15000
	EN	19780	1153	93484	5850	497230	
FR-EN	FR	19661	903	105998	4431	528665	15000
	EN	19993	1208	115722	6161	576543	
Real-World Data							
DOREMUS	G^1	2057	19	5057	3	1775	238
	G^2	1889	20	4659	4	884	
AGROLD	G^1	96117	7	21029	6	28895	11555
	G^2	51488	4	139546	12	225060	

Table 1: Details of the datasets. Rel., Rel.Triples, Attr., Attr.Triples, and Pairs represent relations, relation triples, attributes, attribute triples, and entity pairs respectively.

5.1.2 Baselines

Methods are classified into three categories based on differences in their embedding modules: translation-based methods, GNN-based methods, and BERT-based methods. We have chosen 7 SOTA EA methods that encompass diverse embedding modules. **Translation-based methods:** TransEdge (Sun et al., 2019), MultiKE (Zhang et al., 2019). **GNN-based methods:** RDGCN (Wu et al., 2019a), NMN (Wu et al., 2020). **BERT-based methods:** SDEA (Zhong et al., 2022), BERT-INT (Tang et al., 2020). We also compare the method in Attr-Int (Yang et al., 2024) that calculates only the overlap of attribute value sets.

Although many recent multi-modal entity alignment methods have been developed, like MCLEA (Lin et al., 2022), MEAformer (Chen et al., 2023a), UMEA (Chen et al., 2023b), we do not compare our approach with these methods due to the lack of images in real-world datasets.

5.1.3 Implement details

For each dataset, we divide the aligned entity pairs into training and test sets with a ratio of 3:7. To cover all data from both source and target KGs in one epoch, the batch size for the source KGs are set to 24, and for AGROLD, it is set to 19.

[3] https://github.com/EnsiyehRaoufi\Create_Input_Data_to_EA_Models

However, due to the significant disparity in data volume between the source KGs and DOREMUS, we expand the DOREMUS training set to six times its original size to ensure a thorough traversal of the source KGs. This is achieved by repeating the original training set six times without introducing new data, and the batch size is set to 1.

5.1.4 Evaluation Metric

To facilitate comparison with previous methods, we adopt ranking-based evaluation metrics for entity alignment, specifically Hits@k and mean reciprocal rank (MRR). Hits@k measures the proportion of correct alignments among the top k matches ($k = 1, 10$). Note that higher Hits@k and MRR indicate better performance. We use H@1 and H@10 to present Hits@1 and Hits@10 in this paper.

Methods		DOREMUS			AGROLD		
Emb.Modules	Names	H@1	H@10	MRR	H@1	H@10	MRR
TransE	TransEdge	0.60	4.19	0.036	0.01	0.02	0.001
	MultiKE	2.70	8.70	-	2.30	5.7	-
GCN	RDGCN	1.2	10.9	-	0.02	0.30	-
	NMN	0.0	4.14	-	0.01	0.12	-
BERT	SDEA	38.69	55.95	0.461	0.01	0.02	0.001
	BERT-INT	47.9	59.28	0.515	21.50	25.03	0.229
None	Attr-Int	48.74	76.47	0.587	14.33	20.36	0.167
BERT	**DAEA**	**77.84**	**88.62**	**0.815**	**27.14**	**34.85**	**0.300**
		↑29.94	↑29.34	↑0.3	↑5.64	↑9.82	↑0.071

Table 2: Entity alignment results on Real-World Data

5.2 Experimental Results

5.2.1 Main Results

The experimental results of DAEA compare to other methods on two real-world datasets are shown in Table 2. It is observable that, compared with other methods, DAEA achieves the best performance. Except for BERT-INT and Attr-Int, the performance of the other compared models is relatively suboptimal. The reason for this phenomenon is attributed to the fact that these models incorporate the neighboring entities when calculating the embeddings of entities, whereas BERT-INT only utilizes entity names and descriptions for embedding representation, and Attr-Int merely computes the overlap of attribute value sets. This suggests that in real-world datasets, the neighboring entities of the corresponding entities are highly heterogeneous, introducing noise when neighbor information is included, thus leading to poor alignment results.

To further validate whether the real-world datasets are highly heterogeneous, we employ the method described in (Yang et al., 2024) to calculate the coverage rate of corresponding entities in the real-world datasets compared to those in the benchmark. Let (e_i^1, e_j^2) be an entity pair, $N(e_i^1)$ and $N(e_j^2)$ be the sets of neighboring entities of e_i^1 and e_j^2 respectively, then the coverage rate $C_{(e_i^1,e_j^2)}$ of the entity pair (e_i^1, e_j^2) is calculated by $C_{(e_i^1,e_j^2)} = |N(e_i^1) \cap N(e_j^2)| / \min\left(|N(e_i^1)|, |N(e_j^2)|\right)$, where $|\cdot|$ represents the size of a set.

As illustrated in Figure 4, it can be seen that the neighbours of the corresponding entities in the real-world datasets are completely different, whereas most corresponding entities in the benchmark have the same neighbors. However, most previous models are based on the assumption that identical entities have similar neighboring entities. As a result, these models experience a significant decline in performance on real-world datasets.

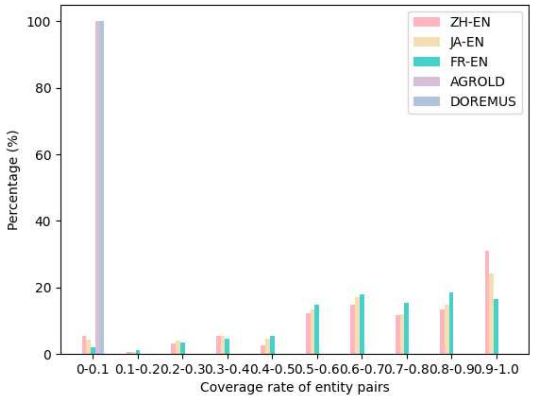

Figure 4: Percentage of coverage rate of entity pairs in each stage of the benchmark datasets and real-world datasets. The x-axis represents the coverage rate of entity pairs, while the y-axis represents the proportion of all benchmark datasets.

5.2.2 Ablation Study

The DAEA model comprises of two main components: multi-source KGs selection and Domain Adaptation. To validate the effectiveness of these components, we conduct ablation studies.

Multi-Source KGs Selection: In the multi-source KGs selection phase, we select source KGs based on the computed distance $\mathcal{D}_{G_sG_t}$ described in Section 4.1, positing that a shorter distance indicates a closer relationship between source and target datasets. As illustrated in Table 3, we quantify the distances between three benchmark datasets (FR-EN, ZH-EN, JA-EN) and two target KGs. Notably, the FR-EN source KGs are closest to the two target KGs and exhibit the best alignment performance. A trend is observed wherein increasing distances between source and target KGs correlated

with decreasing results in entity alignment. This demonstrates the effectiveness of our multi-source KGs selection strategy.

Methods	DOREMUS				AGROLD			
	$\mathcal{D}_{G_sG_t}$	H@1	H@10	MRR	$\mathcal{D}_{G_sG_t}$	H@1	H@10	MRR
FR-EN	**67.51**	**77.84**	**88.62**	**0.815**	**77.64**	**27.14**	**34.85**	**0.300**
ZH-EN	90.72	71.25	85.03	0.756	91.85	22.65	29.26	0.292
JA-EN	105.71	70.65	83.83	0.740	104.55	20.25	28.19	0.231

Table 3: Entity alignment results on Real-World Data with different source KGs.

Domain Adaptation: During the domain adaptation phase, the training process involves the computation of the domain adaptive loss between positive and negative examples from the source KGs and target KGs. To validate the efficacy of the domain adaptive loss and to assess the individual impacts of positive and negative examples, we conduct various experiments. The results are presented in Table 4, where 'DA' denotes domain adaptive loss, 'P' indicates using only positive examples, and 'N' represents using only negative examples. Table 4 illustrates that without domain adaptive loss results in a notable decrease in performance on DOREMUS, with a less significant decline observed on AGROLD. This variation in outcomes can be attributed to the differing sizes of the datasets; AGROLD possesses a considerably larger data volume compared to the source KGs, whereas DOREMUS has a smaller data set. We believe that smaller datasets exhibit simpler data distributions, while larger datasets feature more complex distributions. Consequently, when performing transfer learning with the same source KGs, smaller datasets align more easily with the source KGs, and experimental results tend to be relatively better.

Additionally, on the DOREMUS dataset, the best H@1 score is achieved when both positive and negative examples are transferred. However, on the AGROLD dataset, using only negative examples yields better results. This indicates that in practical transfer learning scenarios, different target datasets cannot be treated uniformly. Instead, the transfer learning approach should be tailored to the specific characteristics and requirements of each target dataset to optimize outcomes.

5.2.3 The impact of dataset size for domain adaptation

To examine the impact of transferring different amounts of data from source KGs to target KGs on entity alignment, we conduct experiments using a fixed training set (30%) in the target KGs, while varying the proportion of entity pairs selected from the source KGs at 30%, 50%, 80%, and 100%. The experimental results are depicted in Figure 5. On the DOREMUS dataset, optimal performance is achieved when 50% of the source data was transferred. Performance do not improve and slightly declines as the transferred data exceeded 50%, suggesting that more source data does not necessarily lead to better alignment. This decline in performance when the source data substantially exceeds the target data may be attributed to an increase in noise within the transferred data.

Conversely, on the AGROLD dataset, the performance improved with an increase in the amount of transferred data. Given the large volume of data in AGROLD, more source data is required for effective transfer learning. In fact, even when 100% of the entity pairs from the source KGs are utilized, the source data volume do not significantly exceed the target data (as opposed to the case for DOREMUS dataset). This suggests that for effective entity alignment through transfer learning, having a larger volume of source data compared to target data is beneficial, as long as the source data maintains a high level of quality and the volume remains within an optimal range.

Methods	DOREMUS			AGROLD		
	H@1	H@10	MRR	H@1	H@10	MRR
DAEA	**77.84**	88.62	**0.815**	27.14	34.85	0.300
-w/o DA	71.86	83.23	0.750	26.24	36.04	0.299
-w P	76.05	**89.82**	0.801	26.87	36.11	0.303
-w N	72.46	84.43	0.762	**27.97**	**37.37**	**0.315**

Table 4: Ablation results. 'w/o' means without and 'w' means with. 'DA' means domain adaptive loss. 'P' indicates using only positive examples, and 'N' represents using only negative examples.

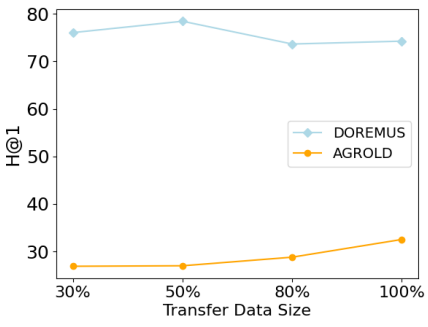

Figure 5: The impact of using training sets of varying sizes from the source KGs for domain adaptation on EA performance.

5.3 Distribution visualization and analysis

To assess the impact of domain adaptation on entity alignment, we compare the DOREMUS entity embeddings before and after the integration of domain adaptation. We visualize the entity embeddings using Principal Components Analysis (PCA) (Maćkiewicz and Ratajczak, 1993), as shown in Figure 6. It can be observed that without domain adaptation (represented in red and blue), the entities are clustered together with almost indistinguishable distances between them, which is a primary cause of suboptimal entity alignment performance. After incorporating domain adaptation (represented in orange and purple), the distances between entities significantly increased, facilitating easier identification of corresponding target entities during alignment and thereby yielding improved entity alignment results.

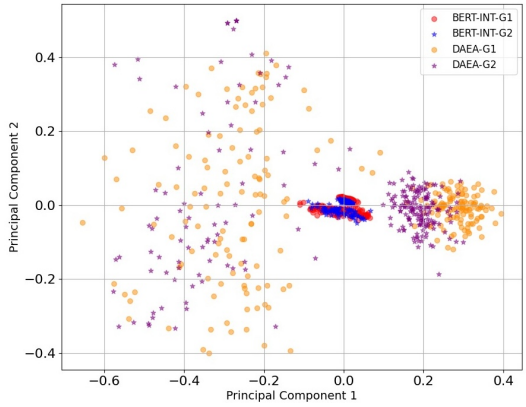

Figure 6: PCA of entity embeddings in DOREMUS. BERT-INT-G1 and BERT-INT-G2 represent the entity embeddings obtained without domain adaptation, DAEA-G1 and DAEA-G2 represent the entity embeddings obtained with domain adaptation.

6 Conclusion

In this paper, we address the issue that current entity alignment models perform well on benchmarks but perform suboptimally on complex real-world datasets. We introduce the DAEA model, which enhances the performance of entity alignment in real-world datasets by leveraging data characteristics from benchmarks through multi-source KGs selection and domain adaptation strategies. Extensive experiments demonstrate that DAEA achieves state-of-the-art performance.

Limitations

While the DAEA model has demonstrated significant improvements in entity alignment performance on real-world datasets, there are still limitations that merit further exploration.

Firstly, the current implementation of DAEA primarily computes the domain adaptive loss on the training sets of the source and target KGs, without extending this transfer learning to the neighboring entities or the entire entity set of the KGs. This constrained scope of domain adaptation may limit the model's ability to fully leverage the structural and semantic richness of the entire KG, potentially affecting the robustness and generalizability of the alignment. Future work will investigate the impact of expanding transfer learning to encompass the complete graph data, aiming to enhance the comprehensiveness and accuracy of entity alignment.

Secondly, the improvements achieved by DAEA are more pronounced on the smaller dataset, DOREMUS, compared to the larger dataset, AGROLD. This disparity suggests that the current domain adaptation strategies may not scale as effectively with increasing data volume. Addressing this challenge, future research will focus on developing new transfer strategies that are better suited to large-scale datasets, thereby improving the model's performance across varying data sizes and complexities.

These limitations highlight the need for ongoing refinement and adaptation of the DAEA model to better address the diverse and dynamic nature of real-world data environments.

Acknowledgement

We sincerely thank the anonymous reviewers for their valuable and insightful feedback, which has greatly contributed to improving the quality of this work. This work is supported by the National Natural Science Foundation of China (62276057), and Sponsored by CAAI-MindSpore Open Fund, developed on OpenI Community. Furthermore, we gratefully acknowledge the additional financial support provided by the China Scholarship Council. Finally, we extend our appreciation to Baskerville for their resources and technical assistance, which played an essential role in the successful completion of this research.

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*.

Yixin Cao, Zhiyuan Liu, Chengjiang Li, Juanzi Li, and Tat-Seng Chua. 2019. Multi-channel graph neural network for entity alignment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1452–1461.

Bo Chen, Jing Zhang, Xiaobin Tang, Hong Chen, and Cuiping Li. 2020. Jarka: Modeling attribute interactions for cross-lingual knowledge alignment. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 845–856. Springer.

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.

Zhuo Chen, Jiaoyan Chen, Wen Zhang, Lingbing Guo, Yin Fang, Yufeng Huang, Yichi Zhang, Yuxia Geng, Jeff Z Pan, Wenting Song, et al. 2023a. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3317–3327.

Zhuo Chen, Lingbing Guo, Yin Fang, Yichi Zhang, Jiaoyan Chen, Jeff Z Pan, Yangning Li, Huajun Chen, and Wen Zhang. 2023b. Rethinking uncertainly missing and ambiguous visual modality in multi-modal entity alignment. In *International Semantic Web Conference*, pages 121–139. Springer.

Kaize Ding, Qinghai Zhou, Hanghang Tong, and Huan Liu. 2021. Few-shot network anomaly detection via cross-network meta-learning. In *Proceedings of the Web Conference 2021*, pages 2448–2456.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 601–610.

Abdullah Elen, Selçuk Baş, and Cemil Közkurt. 2022. An adaptive gaussian kernel for support vector machine. *Arabian Journal for Science and Engineering*, 47(8):10579–10588.

Chenbo Fu, Yongli Zheng, Yi Liu, Qi Xuan, and Guanrong Chen. 2020. Nes-tl: Network embedding similarity-based transfer learning. *IEEE Transactions on Network Science and Engineering*, 7(3):1607–1618.

Bent Fuglede and Flemming Topsoe. 2004. Jensen-shannon divergence and hilbert space embedding. In *International symposium onInformation theory, 2004. ISIT 2004. Proceedings.*, page 31. IEEE.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Hui He, Hongwei Yang, Weizhe Zhang, Yan Wang, Zhaonian Zou, and Tao Li. 2023. Msds: A novel framework for multi-source data selection based cross-network node classification. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12799–12813.

Xuhui Jiang, Chengjin Xu, Yinghan Shen, Yuanzhuo Wang, Fenglong Su, Zhichao Shi, Fei Sun, Zixuan Li, Jian Guo, and Huawei Shen. 2024. Toward practical entity alignment method design: Insights from new highly heterogeneous knowledge graph datasets. In *Proceedings of the ACM on Web Conference 2024*, pages 2325–2336.

Di Jin, Bunyamin Sisman, Hao Wei, Xin Luna Dong, and Danai Koutra. 2021. Deep transfer learning for multi-source entity linkage via domain adaptation. *arXiv preprint arXiv:2110.14509*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks.

Chengjiang Li, Yixin Cao, Lei Hou, Jiaxin Shi, Juanzi Li, and Tat-Seng Chua. 2019. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. In *EMNLP-IJCNLP*, pages 2723–2732.

Zhenxi Lin, Ziheng Zhang, Meng Wang, Yinghui Shi, Xian Wu, and Yefeng Zheng. 2022. Multi-modal contrastive representation learning for entity alignment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2572–2584.

Pasquale Lisena, Manel Achichi, Pierre Choffé, Cécile Cecconi, Konstantin Todorov, Bernard Jacquemin, and Raphaël Troncy. 2018. Improving (re-) usability of musical datasets: An overview of the doremus project. *Bibliothek Forschung und Praxis*, 42(2):194–205.

Andrzej Maćkiewicz and Waldemar Ratajczak. 1993. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342.

Xin Mao, Wenting Wang, Huimin Xu, Man Lan, and Yuanbin Wu. 2020. Mraea: an efficient and robust entity alignment approach for cross-lingual knowledge graph. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 420–428.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. 2018. Multi-adversarial domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543.

Ensiyeh Raoufi, Bill Gates Happi Happi, Pierre Larmande, François Scharffe, and Konstantin Todorov. An analysis of the performance of representation learning methods for entity alignment: Benchmark vs. real-world data.

Xiaofei Shi and Yanghua Xiao. 2019. Modeling multi-mapping relations for precise cross-lingual entity alignment. In *EMNLP-IJCNLP*, pages 813–822.

Baochen Sun and Kate Saenko. 2016. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer.

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 *IJCAI*, volume 18, pages 4396–4402.

Zequn Sun, Jiacheng Huang, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Transedge: Translating relation-contextualized embeddings for knowledge graphs. In *International Semantic Web Conference*, pages 612–629. Springer.

Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. 2020. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 222–229.

Xiaobin Tang, Jing Zhang, Bo Chen, Yang Yang, Hong Chen, and Cuiping Li. 2020. Bert-int: A bert-based interaction model for knowledge graph alignment. In *IJCAI*, pages 3174–3180.

Xiaobin Tian, Zequn Sun, and Wei Hu. 2024. Generating explanations to understand and repair embedding-based entity alignment. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 2205–2217. IEEE.

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 Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. 2017. Graph attention networks. *stat*, 1050(20):10–48550.

Aravind Venkatesan, Gildas Tagny Ngompé, Nordine El Hassouni, Imene Chentli, Valentin Guignon, Clement Jonquet, Manuel Ruiz, and Pierre Larmande. 2018. Agronomic linked data (agrold): A knowledge-based system to enable integrative biology in agronomy. *PLoS One*, 13(11):e0198270.

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.

Zhichun Wang, Jinjian Yang, and Xiaoju Ye. 2020. Knowledge graph alignment with entity-pair embedding. In *EMNLP*, pages 1672–1680.

Y Wu, X Liu, Y Feng, Z Wang, R Yan, and D Zhao. 2019a. 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.

Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao. 2019b. Jointly learning entity and relation representations for entity alignment. In *EMNLP-IJCNLP*, pages 240–249.

Yuting Wu, Xiao Liu, Yansong Feng, Zheng Wang, and Dongyan Zhao. 2020. Neighborhood matching network for entity alignment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6477–6487.

Zhiwen Xie, Runjie Zhu, Kunsong Zhao, Jin Liu, Guangyou Zhou, and Jimmy Xiangji Huang. 2021. Dual gated graph attention networks with dynamic iterative training for cross-lingual entity alignment. *ACM Transactions on Information Systems (TOIS)*, 40(3):1–30.

Zhiwen Xie, Runjie Zhu, Kunsong Zhao, Jin Liu, Guangyou Zhou, and Xiangji Huang. 2020. A contextual alignment enhanced cross graph attention network for cross-lingual entity alignment. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5918–5928.

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.

Zhihuan Yan, Rong Peng, Yaqian Wang, and Weidong Li. 2020. Ctea: Context and topic enhanced entity alignment for knowledge graphs. *Neurocomputing*, 410:419–431.

Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. 2019. Aligning cross-lingual entities with multi-aspect information. In *EMNLP-IJCNLP*, pages 4431–4441.

Linyan Yang, Jingwei Cheng, Chuanhao Xu, Xihao Wang, Jiayi Li, and Fu Zhang. 2024. Attr-int: A simple and effective entity alignment framework for heterogeneous knowledge graphs. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6315–6319. IEEE.

Qingheng Zhang, Zequn Sun, Wei Hu, Muhao Chen, Lingbing Guo, and Yuzhong Qu. 2019. Multi-view knowledge graph embedding for entity alignment.

Ziyi Zhang, Luyi Bai, and Lin Zhu. 2024. Ts-align: A temporal similarity-aware entity alignment model for temporal knowledge graphs. *Information Fusion*, 112:102581.

Ziyue Zhong, Meihui Zhang, Ju Fan, and Chenxiao Dou. 2022. Semantics driven embedding learning for effective entity alignment. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 2127–2140. IEEE.

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.