A Benchmarking Study of Embedding-based Entity Alignment for Knowledge Graphs

Zequn Sun †, Qingheng Zhang †, Wei Hu †* , Chengming Wang †,
Muhao Chen ‡, Farahnaz Akrami §, Chengkai Li §

† State Key Laboratory for Novel Software Technology, Nanjing University, China
‡ Department of Computer Science, University of California, Los Angeles, USA
§ Department of Computer Science and Engineering, University of Texas at Arlington, USA

{zqsun, qhzhang, cmwang}.nju@gmail.com, whu@nju.edu.cn,
muhaochen@ucla.edu, farahnaz.akrami@mavs.uta.edu, cli@uta.edu

VLDB2020

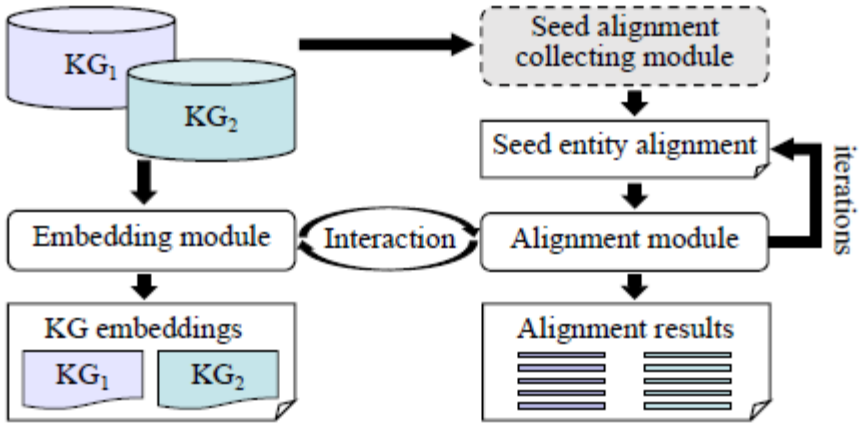

Figure 1: Framework of embedding-based entity alignment

The goal is to identify entities from different KGs that refer to the same entity. There are two typical combination paradigms for module interaction:
(i) the embedding module encodes the two KGs in two independent embedding spaces, meanwhile the alignment module uses seed alignment to learn a mapping between them; or (ii) the alignment module guides the embedding module to represent the two KGs into one unified space by forcing the aligned entities in seed alignment to hold very similar embeddings. Finally, entity similarities are measured by the learned embeddings.

Knowledge graphs (KGs) store facts as triples in the form of (subject entity, relation, object entity) or (subject entity, attribute, literal value).
relation embedding and attribute embedding

Besides, to overcome the shortage of seed entity alignment, several approaches deploy semi-supervised learning to iteratively augment new alignment.

Overview

Sect. 1 INTRODUCTION

Sect. 2 PRELIMINARIES: A comprehensive survey on approaches for embedding-based entity alignment

Sect. 3 Benchmark datasets: propose a new sampling algorithm

Sect. 4 Open-source library: develop an open-source library OpenEA using Python and TensorFlow.

Sect. 5 Comprehensive comparison and analysis

Sect. 6 Exploratory experiments: 1. geometric properties of entity embeddings 2. KG embedding models which have not been exploited for entity alignment 3. Comparison to Conventional Approaches

Sect. 7 Future research directions

Literature Review

Existing KG embedding models can be generally divided into three categories:
(i) translational models, e.g. TransE, TransH, TransR and TransD
(ii) semantic matching models, e.g. DistMult , ComplEx , HolE , SimplE , RotatE and TuckER
(iii) deep models, e.g. ProjE,ConvE, R-GCN, KBGAN and DSKG

Embedding-based Entity Alignment :the translational models & graph convolutional networks (GCNs)

Datasets & evaluation metrics. DBP15K and DWY100K are two benchmark datasets for entity alignment in KGs.

Three metrics are widely used in evaluation: (i) proportion of correct links in the top-m ranked results (called Hits@m, for example, m = 1), (ii) mean rank (MR) of correct links, and (iii) mean reciprocal rank (MRR).

Current datasets contain much more high-degree entities (i.e., entities connected with many other entities, which are relatively easy for entity alignment) than real world KGs do. As a result, many approaches may exhibit good performance on these biased datasets.

Additionally, these datasets only focus on one aspect of heterogeneity, e.g., multilingualism, while overlook other aspects, e.g, different schemata and scales. This brings difficulties in understanding the generalization and robustness of embedding-based entity alignment.

Categorization of Techniques

1. Embedding Module 2. Alignment Module 3. Interaction Module

Embedding Module: 1.Triple-based embedding 2. Path-based embedding 3. Neighborhood-based embedding 4. Attribute correlation embedding 5. Literal embedding

Relation embedding is employed by all existing approaches.

- Triple-based embedding captures the local semantics of relation triples. $\phi(e_1, r_1, e_2) = \| \mathbf{e}_1 + \mathbf{r}_1 - \mathbf{e}_2 \|,$

- Path-based embedding: exploit the long-term dependency of relations spanning over relation paths

  IPTransE models relation paths by inferring the equivalence between a direct relation and a multi-hop path. Assume that there is a direct relation r_3 from e_1 to e_3. IPTransE expects the embedding of r_3 to be similar to the path embedding, which is encoded as a combination of its constituent relation embeddings:

  $\mathbf{r}^* = \text{comb}(\mathbf{r}_1, \mathbf{r}_2),$ minimized $\| \mathbf{r}^* - \mathbf{r}_3 \|$

- Neighborhood-based embedding $\mathbf{H}^{(i+1)} = \sigma(\hat{\mathbf{D}}^{-\frac{1}{2}} \hat{\mathbf{A}} \hat{\mathbf{D}}^{-\frac{1}{2}} \mathbf{H}^{(i)} \mathbf{W}),$

Attribute embedding

There are two ways for attribute embedding:

Attribute correlation embedding: consider the correlations among attributes

Attributes are regarded as correlated if they are frequently used together to describe an entity. For example, longitude is highly correlated with latitude as they often form a coordinate.

JAPE exploits such correlations for entity alignment, based on the assumption that similar entities should have similar correlated attributes.

$$\Pr(a_1, a_2) = \text{sigmoid}(\mathbf{a}_1 \cdot \mathbf{a}_2),$$

Attribute embeddings can be learned by maximizing the probability over all attribute pairs

Literal embedding: introduces literal values to attribute embedding.

Let $v = (c_1, c_2, ..., c_n)$ be a literal with n characters, where c_i $(1 \leq i \leq n)$ is the i^{th} character. AttrE embeds v as

$$\mathbf{v} = \text{comb}(\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_n).$$

With this representation, literals are treated as entities and the relation embedding models like TransE can be used to learn from attribute triples. However, the character-based literal embedding may fail in cross-lingual settings.

Alignment Module

1. Pick a distance metric
2. Design alignment inference strategy.

Distance metrics: Cosine, Euclidean and Manhattan distances are three widely-used metrics. In high-dimensional spaces, a few vectors (called hubs) may repeatedly occur as the k-nearest neighbors of others, the so-called hubness problem.

Alignment inference strategies:

1. Greedy search
2. Collective search: find a global optimal alignment that minimizes $\sum_{(e_1,e_2)\in\mathcal{S}_{\mathcal{KG}_1,\mathcal{KG}_2}} \pi(\mathbf{e}_1,\mathbf{e}_2)$ $O(N^3)$

Another solution is the stable marriage algorithm. The alignment between E_1 and E_2 satisfies a stable marriage if there does not exist a pair of entities that both prefer each other than their current aligned ones. Its solution takes $O(N^2)$ time

Interaction Mode

1. Combination modes.

- Embedding space transformation: embeds two KGs in different embedding spaces and learns a transformation matrix M between the two spaces. $\mathbf{M}\mathbf{e}_1 \approx \mathbf{e}_2$
- Embedding space calibration: minimizes $\| \mathbf{e}_1 - \mathbf{e}_2 \|$ for each $(e_1, e_2) \in \mathcal{S}'_{\mathcal{KG}_1, \mathcal{KG}_2}$
- Parameter sharing: directly configures e1 = e2
- Parameter swapping: swap seed entities in their triples to generate extra triples as supervision.

2. Learning strategies.

- Supervised learning
- Semi-supervised learning:
  - self-training： iteratively propose new alignment to augment seed alignment.
  - co-training：combine two models learned from disjoint entity features and alternately enhances the alignment learning of each other.
- Unsupervised learning

Sect. 3 Benchmark datasets

Iterative Degree-based Sampling

Iterative degree-based sampling (IDS) algorithm, which simultaneously deletes entities in two source KGs with reference alignment until achieving the desired size, meanwhile keeping a similar degree distribution of each sampled dataset as the source KG.

pling procedure. During iterations, the proportion of entities having degree x in the current dataset, denoted by $P(x)$, cannot always equal the original proportion $Q(x)$. We adjust the entity size to be deleted by $dsize(x,\mu) = \mu(1+P(x)-Q(x))$, where μ is the base step size (see Line 7). Moreover, we prefer not to delete entities having a big influence on the overall degree distribution, such as the ones of high degree. To achieve this, we leverage the PageRank value for measuring the probability of an entity to be deleted (Line 8).

Algorithm 1: Iterative degree-based sampling (IDS)

Input: $\mathcal{KG}_1, \mathcal{KG}_2$, reference alignment $\mathcal{S}_{ref}$, entity size N, hyper-parameters μ, ϵ

`// only retain entities in reference alignment`

1 Filter $\mathcal{KG}_1, \mathcal{KG}_2$ by $\mathcal{S}_{ref}$;
2 Get degree distributions Q_1, Q_2 for $\mathcal{KG}_1, \mathcal{KG}_2$, resp.;
3 **do** `// if fails, run it again`
4 Initialize datasets $\mathcal{DS}_1, \mathcal{DS}_2$ from $\mathcal{KG}_1, \mathcal{KG}_2$, resp.;
5 **while** $|\mathcal{DS}_1| > N$ && $|\mathcal{DS}_2| > N$ **do**
6 **for** $\mathcal{DS}_j$ $(j = 1, 2)$ **do**
7 Get $dsize_j(x, \mu)$ for each degree x;
8 Get entity deletion probability by PageRank;
9 Delete $dsize_j(x, \mu)$ entities w.r.t. probabilities;
10 Filter $\mathcal{DS}_1, \mathcal{DS}_2$ by $\mathcal{S}_{ref}$; update $\mathcal{S}_{ref}$ accordingly;
11 Get degree distributions P_1, P_2 for $\mathcal{DS}_1, \mathcal{DS}_2$, resp.;
12 **while** $JS(Q_1, P_1) > \epsilon \parallel JS(Q_2, P_2) > \epsilon$;
13 **return** $\mathcal{DS}_1, \mathcal{DS}_2, \mathcal{S}_{ref}$;

The difference of two degree distributions

$$JS(Q,P) = \frac{1}{2}\sum_{x=1}^{n}\left(Q(x)\log\frac{Q(x)}{M(x)} + P(x)\log\frac{P(x)}{M(x)}\right)$$

where $Q(x)$ and $P(x)$ denote the proportions of entities with degree x $(x = 1 \ldots n)$ in Q, P, respectively, and $M = \frac{Q+P}{2}$.

A small JS divergence between Q and P reveals that they have similar degree distributions.

Dataset Evaluation

We generate two versions of datasets for each pair of source KGs. V1 is gained by directly using the IDS algorithm. For V2, we first randomly delete entities with low degrees (d≤5) in the source KG to make the average degree doubled, and then execute IDS to t the new KG. As a result, V2 is twice denser than V1 and more similar to existing datasets

- *Random alignment sampling (RAS)* first randomly selects a fixed size (e.g., 15K) of entity alignment between two KGs, and then extracts the relation triples whose head and tail entities are both in the sampled entities.
- *PageRank-based sampling (PRS)* first samples entities from one KG based on the PageRank scores (entities not involved in any alignment are discarded), and then extracts these entities' counterparts from the other KG.

Table 2: Dataset statistics

Datasets	KGs	15K (V1)				15K (V2)				100K (V1)				100K (V2)			
		#Rel.	#Att.	#Rel tr.	#Att tr.	#Rel.	#Att.	#Rel tr.	#Att tr.	#Rel.	#Att.	#Rel tr.	#Att tr.	#Rel.	#Att.	#Rel tr.	#Att tr.
EN-FR	EN	267	308	47,334	73,121	193	189	96,318	66,899	400	466	309,607	497,729	379	364	649,902	503,922
	FR	210	404	40,864	67,167	166	221	80,112	68,779	300	519	258,285	426,672	287	468	561,391	431,379
EN-DE	EN	215	286	47,676	83,755	169	171	84,867	81,988	381	451	335,359	552,750	323	326	622,588	560,247
	DE	131	194	50,419	156,150	96	116	92,632	186,335	196	252	336,240	716,615	170	189	629,395	793,710
D-W	DB	248	342	38,265	68,258	167	175	73,983	66,813	413	493	293,990	451,011	318	328	616,457	467,103
	WD	169	649	42,746	138,246	121	457	83,365	175,686	261	874	251,708	687,860	239	760	588,203	878,219
D-Y	DB	165	257	30,291	71,716	72	90	68,063	65,100	287	379	294,188	523,062	230	277	576,547	547,026
	YG	28	35	26,638	132,114	21	20	60,970	131,151	32	38	400,518	749,787	31	36	865,265	855,161

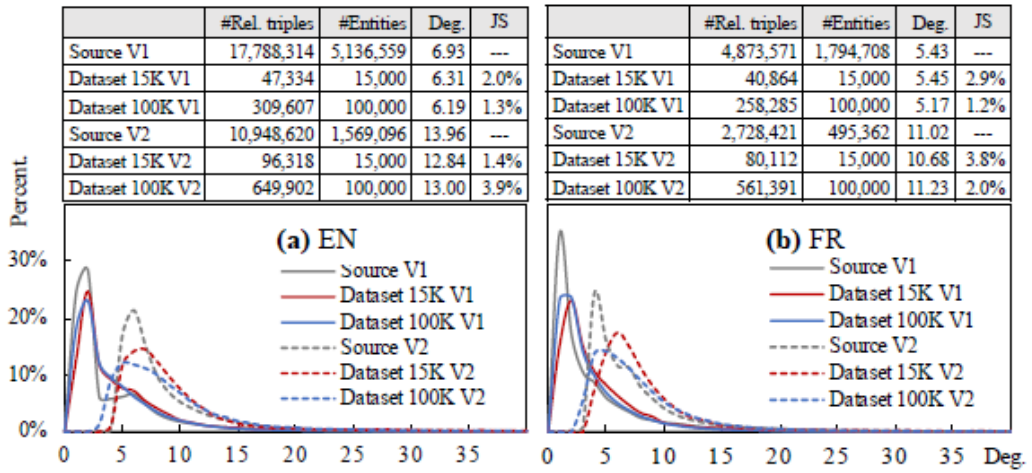

Figure 3: Degree distributions and average degrees of our sampled datasets EN-FR-15K (V1, V2) and EN-FR-100K (V1, V2), compared with DBpedia (the source KG)

Table 3: Comparison of the EN-FR-15K (V1) datasets generated by RAS, PRS and IDS

Datasets	KGs	#Alignment	Deg.	JS	Isolates	Cluster coef.
DBpedia	EN	525,807	6.39	–	0	0.342
	FR		5.43	–	0	0.080
RAS	EN	15,000	0.27	14.5%	85.5%	0.002
	FR		0.17	12.1%	90.1%	0.001
PRS	EN	15,000	1.20	7.3%	68.9%	0.025
	FR		0.63	9.3%	69.4%	0.015
IDS	EN	15,000	6.31	2.0%	0	0.233
	FR		5.45	2.9%	0	0.190

Sect. 4 Open-source library

Sect. 5 Comprehensive comparison and analysis

Table 5: Cross-validation results of current representative approaches on the 15K and 100K datasets

		15K (V1)			15K (V2)			100K (V1)			100K (V2)		
		Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR
EN-FR	MTransE	.247 ± .006	.467 ± .009	.351 ± .007	.240 ± .005	.436 ± .007	.336 ± .005	.138 ± .002	.261 ± .004	.202 ± .002	.090 ± .003	.174 ± .003	.135 ± .003
	IPTransE	.169 ± .013	.320 ± .025	.243 ± .019	.236 ± .012	.449 ± .021	.339 ± .016	.158 ± .004	.277 ± .008	.219 ± .006	.234 ± .007	.431 ± .015	.329 ± .010
	JAPE	.262 ± .006	.497 ± .010	.372 ± .007	.292 ± .009	.524 ± .006	.402 ± .007	.165 ± .002	.310 ± .002	.240 ± .002	.125 ± .003	.239 ± .005	.183 ± .004
	KDCoE	.581 ± .004	.680 ± .004	.628 ± .003	.730 ± .007	.837 ± .006	.778 ± .005	.482 ± .005	.515 ± .006	.499 ± .005	.611 ± .012	.653 ± .015	.632 ± .014
	BootEA	.507 ± .010	.718 ± .012	.603 ± .011	.660 ± .006	.850 ± .005	.745 ± .005	.389 ± .004	.561 ± .004	.474 ± .004	.640 ± .001	.806 ± .001	.716 ± .000
	GCNAlign	.338 ± .002	.589 ± .009	.451 ± .005	.414 ± .005	.698 ± .007	.542 ± .005	.230 ± .002	.412 ± .004	.319 ± .003	.257 ± .002	.455 ± .003	.351 ± .002
	AttrE	.481 ± .010	.671 ± .009	.569 ± .010	.535 ± .015	.746 ± .014	.631 ± .014	.403 ± .019	.572 ± .019	.483 ± .019	.466 ± .011	.644 ± .012	.549 ± .011
	IMUSE	.569 ± .006	.717 ± .010	.638 ± .008	.607 ± .013	.760 ± .014	.678 ± .013	.439 ± .002	.546 ± .004	.492 ± .003	.461 ± .003	.605 ± .005	.529 ± .004
	SEA	.280 ± .015	.530 ± .026	.397 ± .019	.360 ± .018	.651 ± .018	.494 ± .017	.225 ± .011	.399 ± .013	.314 ± .012	.297 ± .002	.500 ± .002	.395 ± .002
	RSN4EA	.393 ± .007	.595 ± .012	.487 ± .009	.579 ± .006	.759 ± .006	.662 ± .006	.293 ± .004	.452 ± .006	.371 ± .004	.495 ± .003	.672 ± .005	.578 ± .004
	MultiKE	.749 ± .004	.819 ± .005	.782 ± .004	.864 ± .007	.909 ± .005	.885 ± .006	.629 ± .002	.680 ± .002	.655 ± .002	.642 ± .003	.696 ± .003	.670 ± .003
	RDGCN	.755 ± .004	.854 ± .003	.800 ± .003	.847 ± .006	.919 ± .004	.880 ± .005	.640 ± .004	.732 ± .004	.683 ± .004	.715 ± .003	.787 ± .002	.748 ± .002
EN-DE	MTransE	.307 ± .007	.518 ± .004	.407 ± .006	.193 ± .016	.352 ± .023	.274 ± .018	.140 ± .003	.264 ± .004	.204 ± .004	.115 ± .003	.215 ± .004	.168 ± .003
	IPTransE	.350 ± .009	.515 ± .012	.43 ± .011	.476 ± .012	.678 ± .011	.571 ± .010	.226 ± .014	.357 ± .019	.292 ± .017	.346 ± .013	.535 ± .016	.437 ± .014
	JAPE	.288 ± .016	.512 ± .018	.394 ± .016	.167 ± .011	.329 ± .015	.250 ± .013	.152 ± .006	.291 ± .009	.223 ± .007	.11 ± .004	.218 ± .006	.167 ± .005
	KDCoE	.529 ± .014	.629 ± .015	.580 ± .014	.649 ± .017	.788 ± .017	.715 ± .016	.506 ± .014	.591 ± .019	.549 ± .016	.651 ± .011	.756 ± .010	.701 ± .011
	BootEA	.675 ± .004	.820 ± .004	.740 ± .004	.833 ± .015	.912 ± .008	.869 ± .012	.518 ± .003	.673 ± .003	.592 ± .003	.739 ± .004	.851 ± .003	.791 ± .004
	GCNAlign	.481 ± .003	.679 ± .005	.571 ± .003	.534 ± .005	.717 ± .005	.618 ± .005	.317 ± .007	.485 ± .008	.399 ± .007	.375 ± .005	.549 ± .006	.457 ± .005
	AttrE	.517 ± .011	.687 ± .013	.597 ± .011	.650 ± .015	.816 ± .008	.726 ± .012	.399 ± .010	.554 ± .012	.473 ± .011	.464 ± .011	.637 ± .010	.546 ± .011
	IMUSE	.580 ± .017	.720 ± .014	.647 ± .015	.674 ± .011	.803 ± .008	.734 ± .010	.421 ± .005	.516 ± .005	.469 ± .005	.457 ± .005	.588 ± .007	.521 ± .006
	SEA	.530 ± .027	.718 ± .026	.617 ± .025	.606 ± .024	.779 ± .018	.687 ± .020	.341 ± .016	.502 ± .017	.421 ± .016	.447 ± .006	.625 ± .006	.532 ± .006
	RSN4EA	.587 ± .001	.752 ± .003	.662 ± .001	.791 ± .009	.890 ± .006	.837 ± .008	.430 ± .002	.57 ± .001	.497 ± .001	.639 ± .001	.763 ± .001	.697 ± .001
	MultiKE	.756 ± .004	.809 ± .003	.782 ± .003	.755 ± .008	.813 ± .008	.784 ± .007	.668 ± .002	.712 ± .002	.690 ± .001	.661 ± .004	.709 ± .004	.686 ± .004
	RDGCN	.830 ± .006	.895 ± .004	.859 ± .005	.833 ± .007	.891 ± .005	.860 ± .006	.722 ± .002	.794 ± .002	.756 ± .002	.766 ± .002	.829 ± .002	.796 ± .002
D-W	MTransE	.259 ± .008	.461 ± .012	.354 ± .008	.271 ± .013	.49 ± .014	.376 ± .013	.210 ± .003	.358 ± .003	.282 ± .003	.148 ± .004	.268 ± .005	.209 ± .005
	IPTransE	.232 ± .012	.38 ± .016	.303 ± .014	.412 ± .007	.623 ± .010	.511 ± .007	.221 ± .004	.352 ± .008	.285 ± .006	.319 ± .017	.516 ± .024	.413 ± .020
	JAPE	.250 ± .007	.457 ± .010	.348 ± .007	.262 ± .013	.484 ± .019	.368 ± .015	.211 ± .004	.369 ± .004	.287 ± .004	.154 ± .004	.287 ± .005	.221 ± .005
	KDCoE	.247 ± .020	.412 ± .029	.325 ± .023	.405 ± .020	.640 ± .019	.515 ± .020	.157 ± .003	.243 ± .007	.199 ± .005	.373 ± .010	.550 ± .014	.458 ± .012
	BootEA	.572 ± .008	.744 ± .007	.649 ± .008	.821 ± .004	.926 ± .003	.867 ± .003	.516 ± .006	.685 ± .006	.594 ± .005	.766 ± .007	.892 ± .005	.822 ± .006
	GCNAlign	.364 ± .009	.580 ± .010	.461 ± .008	.506 ± .006	.743 ± .005	.612 ± .005	.324 ± .002	.507 ± .004	.409 ± .003	.353 ± .004	.559 ± .006	.449 ± .004
	AttrE	.299 ± .004	.467 ± .003	.381 ± .003	.489 ± .016	.695 ± .016	.585 ± .015	.209 ± .008	.335 ± .011	.273 ± .009	.301 ± .015	.475 ± .018	.386 ± .016
	IMUSE	.327 ± .016	.523 ± .024	.419 ± .019	.581 ± .016	.778 ± .011	.671 ± .014	.276 ± .010	.437 ± .016	.355 ± .013	.431 ± .011	.631 ± .013	.525 ± .012
	SEA	.360 ± .012	.572 ± .015	.458 ± .013	.567 ± .008	.770 ± .007	.660 ± .008	.291 ± .012	.470 ± .014	.378 ± .013	.382 ± .003	.585 ± .003	.479 ± .002
	RSN4EA	.441 ± .008	.615 ± .007	.521 ± .007	.723 ± .007	.854 ± .006	.782 ± .006	.384 ± .004	.533 ± .006	.454 ± .005	.634 ± .004	.776 ± .002	.699 ± .003
	MultiKE	.411 ± .010	.521 ± .017	.468 ± .012	.495 ± .010	.646 ± .016	.569 ± .011	.290 ± .006	.357 ± .011	.326 ± .009	.319 ± .001	.396 ± .004	.360 ± .002
	RDGCN	.515 ± .008	.669 ± .006	.584 ± .007	.623 ± .006	.757 ± .004	.684 ± .005	.362 ± .002	.485 ± .002	.420 ± .002	.421 ± .001	.528 ± .002	.473 ± .002
D-Y	MTransE	.463 ± .013	.675 ± .011	.559 ± .012	.443 ± .017	.635 ± .013	.533 ± .015	.244 ± .004	.414 ± .006	.328 ± .005	.100 ± .003	.195 ± .005	.152 ± .004
	IPTransE	.313 ± .009	.456 ± .015	.378 ± .011	.752 ± .018	.873 ± .013	.808 ± .015	.396 ± .014	.558 ± .018	.474 ± .015	.456 ± .016	.620 ± .017	.534 ± .017
	JAPE	.469 ± .009	.687 ± .011	.567 ± .009	.345 ± .010	.546 ± .013	.440 ± .011	.287 ± .007	.474 ± .008	.379 ± .007	.127 ± .004	.244 ± .006	.189 ± .005
	KDCoE	.661 ± .013	.764 ± .036	.710 ± .021	.895 ± .013	.974 ± .003	.932 ± .008	.565 ± .001	.646 ± .002	.605 ± .002	.540 ± .001	.621 ± .002	.581 ± .001
	BootEA	.739 ± .014	.849 ± .010	.788 ± .012	.958 ± .001	.984 ± .001	.969 ± .001	.703 ± .004	.827 ± .003	.761 ± .003	.886 ± .003	.944 ± .002	.912 ± .002
	GCNAlign	.465 ± .012	.626 ± .011	.536 ± .011	.875 ± .005	.948 ± .004	.907 ± .004	.528 ± .003	.695 ± .005	.605 ± .004	.620 ± .006	.779 ± .006	.693 ± .006
	AttrE	.668 ± .012	.803 ± .009	.731 ± .010	.914 ± .015	.97 ± .007	.939 ± .012	.678 ± .017	.81 ± .012	.739 ± .015	.720 ± .01	.846 ± .007	.778 ± .008
	IMUSE	.392 ± .013	.571 ± .023	.473 ± .017	.899 ± .011	.949 ± .007	.922 ± .009	.536 ± .018	.700 ± .019	.613 ± .018	.629 ± .011	.774 ± .010	.696 ± .010
	SEA	.500 ± .011	.706 ± .012	.591 ± .012	.899 ± .005	.950 ± .003	.923 ± .004	.490 ± .042	.677 ± .040	.578 ± .040	.526 ± .021	.687 ± .021	.603 ± .021
	RSN4EA	.514 ± .003	.655 ± .004	.580 ± .003	.933 ± .003	.974 ± .001	.951 ± .002	.620 ± .002	.769 ± .003	.688 ± .002	.841 ± .003	.922 ± .001	.877 ± .002
	MultiKE	.903 ± .004	.939 ± .003	.920 ± .003	.856 ± .004	.908 ± .002	.881 ± .003	.884 ± .005	.920 ± .004	.901 ± .005	.853 ± .003	.896 ± .003	.874 ± .003
	RDGCN	.931 ± .004	.969 ± .003	.949 ± .003	.936 ± .003	.966 ± .001	.950 ± .002	.897 ± .001	.950 ± .001	.921 ± .001	.911 ± .002	.949 ± .002	.928 ± .002

Means $_{\pm \text{ stds.}}$ are shown. Top-3 results on each dataset are marked in red, blue and cyan, respectively. The same to the following.

		15K (V1)			15K (V2)			100K (V1)			100K (V2)		
		Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR
EN-FR	MTransE	.247 ± .006	.467 ± .009	.351 ± .007	.240 ± .005	.436 ± .007	.336 ± .005	.138 ± .002	.261 ± .004	.202 ± .002	.090 ± .003	.174 ± .003	.135 ± .003
	IPTransE	.169 ± .013	.320 ± .025	.243 ± .019	.236 ± .012	.449 ± .021	.339 ± .016	.158 ± .004	.277 ± .008	.219 ± .006	.234 ± .007	.431 ± .015	.329 ± .010
	JAPE	.262 ± .006	.497 ± .010	.372 ± .007	.292 ± .009	.524 ± .006	.402 ± .007	.165 ± .002	.310 ± .002	.240 ± .002	.125 ± .003	.239 ± .005	.183 ± .004
	KDCoE	.581 ± .004	.680 ± .004	.628 ± .003	.730 ± .007	.837 ± .006	.778 ± .005	.482 ± .005	.515 ± .006	.499 ± .005	.611 ± .012	.653 ± .015	.632 ± .014
	BootEA	.507 ± .010	.718 ± .012	.603 ± .011	.660 ± .006	.850 ± .005	.745 ± .005	.389 ± .004	.561 ± .004	.474 ± .004	.640 ± .001	.806 ± .001	.716 ± .000
	GCNAlign	.338 ± .002	.589 ± .009	.451 ± .005	.414 ± .005	.698 ± .007	.542 ± .005	.230 ± .002	.412 ± .004	.319 ± .003	.257 ± .002	.455 ± .003	.351 ± .002
	AttrE	.481 ± .010	.671 ± .009	.569 ± .010	.535 ± .015	.746 ± .014	.631 ± .014	.403 ± .019	.572 ± .019	.483 ± .019	.466 ± .011	.644 ± .012	.549 ± .011
	IMUSE	.569 ± .006	.717 ± .010	.638 ± .008	.607 ± .013	.760 ± .014	.678 ± .013	.439 ± .002	.546 ± .004	.492 ± .003	.461 ± .003	.605 ± .005	.529 ± .004
	SEA	.280 ± .015	.530 ± .026	.397 ± .019	.360 ± .018	.651 ± .018	.494 ± .017	.225 ± .011	.399 ± .013	.314 ± .012	.297 ± .002	.500 ± .002	.395 ± .002
	RSN4EA	.393 ± .007	.595 ± .012	.487 ± .009	.579 ± .006	.759 ± .006	.662 ± .006	.293 ± .004	.452 ± .006	.371 ± .004	.495 ± .003	.672 ± .005	.578 ± .004
	MultiKE	.749 ± .004	.819 ± .005	.782 ± .004	.864 ± .007	.909 ± .005	.885 ± .006	.629 ± .002	.680 ± .002	.655 ± .002	.642 ± .003	.696 ± .003	.670 ± .003
	RDGCN	.755 ± .004	.854 ± .003	.800 ± .003	.847 ± .006	.919 ± .004	.880 ± .005	.640 ± .004	.732 ± .004	.683 ± .004	.715 ± .003	.787 ± .002	.748 ± .002

Sparse datasets (V1) vs. dense datasets (V2).

Most relation-based approaches perform better on the dense datasets than on the sparse ones, e.g., IPTransE, BootEA, SEA and RSN4EA.

For the approaches considering attribute triples, KDCoE, GCNAlign, AttrE, IMUSE and RDGCN also perform better on the dense datasets, indicating that the relation embeddings still make contributions.

Differently, MultiKE relies on multiple "views“ of features, which make it relatively insensitive to the relation changes. Interestingly, we also see that the performance of two relation-based approaches, MTransE and JAPE, drops on some dense datasets. We believe that this is because they are based on TransE, which has deciency in handling multi-mapping relations in the dense datasets.

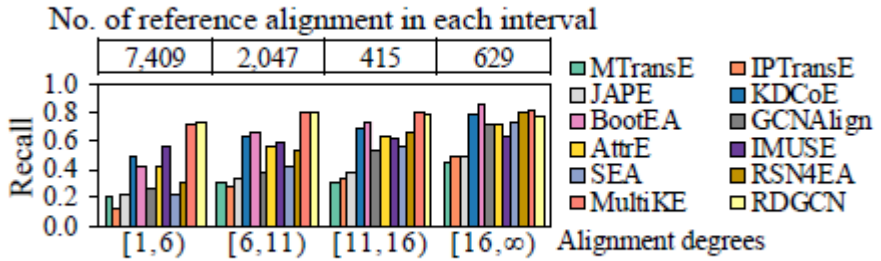

Figure 5: Recall w.r.t. alignment deg. on EN-FR-15K (V1)

We find that all the relation-based approaches run better in aligning entities with rich relation triples while their results decline on long-tail entities,

15K datasets vs. 100K datasets.

We observe that all the approaches perform better on the 15K datasets than on the 100K datasets, except D-Y, because the 100K datasets have more complex structures, causing more difficulties for embedding-based approaches to capture entity proximity.

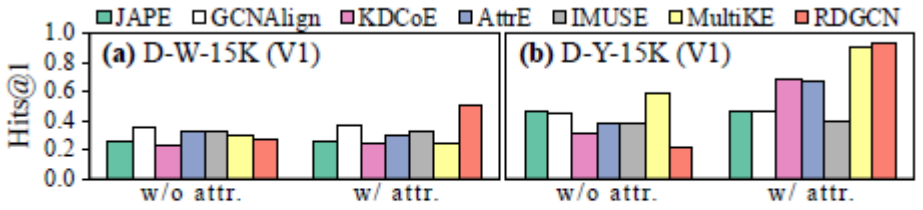

Figure 6: Hits@1 results of JAPE, GCNAlign, KDCoE, AttrE, IMUSE, MultiKE, RDGCN and their degraded variants without attribute embedding

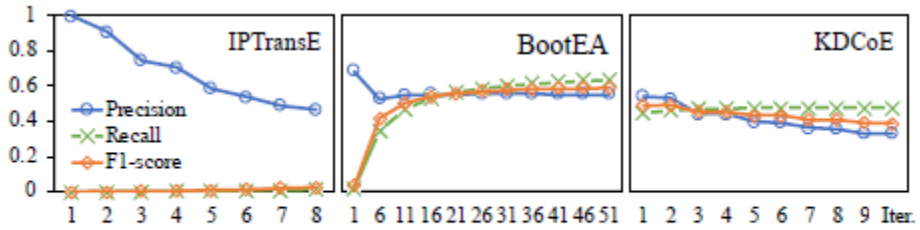

Figure 7: Precision, recall and F1-score of augmented alignment during iterations on EN-FR-100K (V1)

Sect. 6 EXPLORATORY EXPERIMENTS

Similarity Distribution

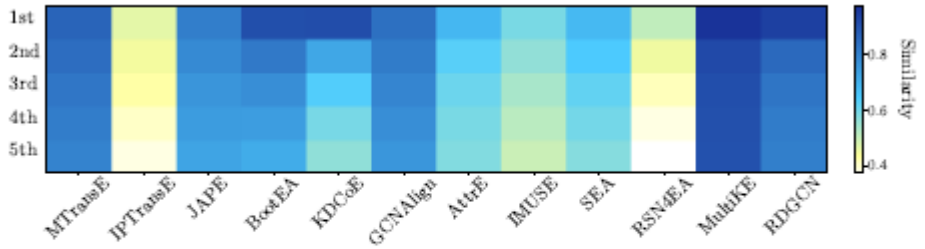

Figure 9: Visualization of the similarities between entities and their top-5 nearest cross-KG neighbors on the D-Y-15K (V1) dataset. The five rows from top to bottom correspond to the similarities from the first to the fifth nearest neighbors, respectively. Darker color indicates larger similarity.

BootEA, KDCoE, MultiKE and RDGCN yield a very high top-1 similarity, while IPTransE and RSN4EA show the opposite. Intuitively, a high top-1 similarity indicates a better quality because it can reflect how confidently the entity embeddings capture the alignment information between two KGs. Most approaches with a high top-1 similarity, such as BootEA, MultiKE and RDGCN, also achieve good performance for entity alignment

The similarity variances between the top-5 nearest neighbors also differ greatly

The ideal similarity distribution for entity alignment is to hold a high top-1 similarity and a large similarity variance.

		15K (V1)			15K (V2)			100K (V1)			100K (V2)		
		Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR
EN-FR	MTransE	.247 ± .006	.467 ± .009	.351 ± .007	.240 ± .005	.436 ± .007	.336 ± .005	.138 ± .002	.261 ± .004	.202 ± .002	.090 ± .003	.174 ± .003	.135 ± .003
	IPTransE	.169 ± .013	.320 ± .025	.243 ± .019	.236 ± .012	.449 ± .021	.339 ± .016	.158 ± .004	.277 ± .008	.219 ± .006	.234 ± .007	.431 ± .015	.329 ± .010
	JAPE	.262 ± .006	.497 ± .010	.372 ± .007	.292 ± .009	.524 ± .006	.402 ± .007	.165 ± .002	.310 ± .002	.240 ± .002	.125 ± .003	.239 ± .005	.183 ± .004
	KDCoE	.581 ± .004	.680 ± .004	.628 ± .003	.730 ± .007	.837 ± .006	.778 ± .005	.482 ± .005	.515 ± .006	.499 ± .005	.611 ± .012	.653 ± .015	.632 ± .014
	BootEA	.507 ± .010	.718 ± .012	.603 ± .011	.660 ± .006	.850 ± .005	.745 ± .005	.389 ± .004	.561 ± .004	.474 ± .004	.640 ± .001	.806 ± .001	.716 ± .000
	GCNAlign	.338 ± .002	.589 ± .009	.451 ± .005	.414 ± .005	.698 ± .007	.542 ± .005	.230 ± .002	.412 ± .004	.319 ± .003	.257 ± .002	.455 ± .003	.351 ± .002
	AttrE	.481 ± .010	.671 ± .009	.569 ± .010	.535 ± .015	.746 ± .014	.631 ± .014	.403 ± .019	.572 ± .019	.483 ± .019	.466 ± .011	.644 ± .012	.549 ± .011
	IMUSE	.569 ± .006	.717 ± .010	.638 ± .008	.607 ± .013	.760 ± .014	.678 ± .013	.439 ± .002	.546 ± .004	.492 ± .003	.461 ± .003	.605 ± .005	.529 ± .004
	SEA	.280 ± .015	.530 ± .026	.397 ± .019	.360 ± .018	.651 ± .018	.494 ± .017	.225 ± .011	.399 ± .013	.314 ± .012	.297 ± .002	.500 ± .002	.395 ± .002
	RSN4EA	.393 ± .007	.595 ± .012	.487 ± .009	.579 ± .006	.759 ± .006	.662 ± .006	.293 ± .004	.452 ± .006	.371 ± .004	.495 ± .003	.672 ± .005	.578 ± .004
	MultiKE	.749 ± .004	.819 ± .005	.782 ± .004	.864 ± .007	.909 ± .005	.885 ± .006	.629 ± .002	.680 ± .002	.655 ± .002	.642 ± .003	.696 ± .003	.670 ± .003
	RDGCN	.755 ± .004	.854 ± .003	.800 ± .003	.847 ± .006	.919 ± .004	.880 ± .005	.640 ± .004	.732 ± .004	.683 ± .004	.715 ± .003	.787 ± .002	.748 ± .002

Hubness and Isolation

We observe that the approaches which yield fewer isolated and hub entities, such as MultiKE and RDGCN, achieve the leading performance of entity alignment, and vice versa.

To resolve the hubness and isolation problem, we explore cross-domain similarity local scaling (CSLS)

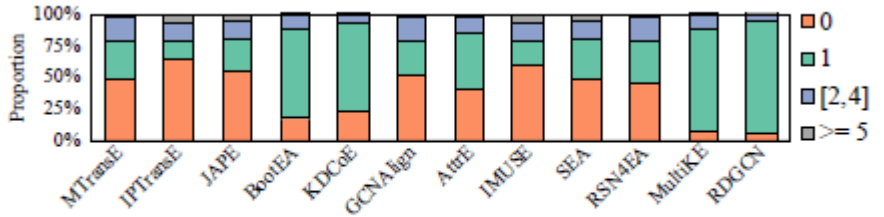

Figure 10: Proportions of target entities appearing 0, 1 times as the nearest neighbors on D-Y-15K (V1)

$$\mathrm{CSLS}(\mathbf{x}_s, \mathbf{x}_t) = 2\cos(\mathbf{x}_s, \mathbf{x}_t) - \psi_t(\mathbf{x}_s) - \psi_s(\mathbf{x}_t),$$

$\psi_t(\mathbf{x}_s)$ denotes the average similarity between the source entity $\mathbf{x}_s$ and its top-k nearest neighbors in the target KG.

		15K (V1)			15K (V2)			100K (V1)			100K (V2)		
		Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR	Hits@1	Hits@5	MRR
EN-FR	MTransE	.247 ± .006	.467 ± .009	.351 ± .007	.240 ± .005	.436 ± .007	.336 ± .005	.138 ± .002	.261 ± .004	.202 ± .002	.090 ± .003	.174 ± .003	.135 ± .003
	IPTransE	.169 ± .013	.320 ± .025	.243 ± .019	.236 ± .012	.449 ± .021	.339 ± .016	.158 ± .004	.277 ± .008	.219 ± .006	.234 ± .007	.431 ± .015	.329 ± .010
	JAPE	.262 ± .006	.497 ± .010	.372 ± .007	.292 ± .009	.524 ± .006	.402 ± .007	.165 ± .002	.310 ± .002	.240 ± .002	.125 ± .003	.239 ± .005	.183 ± .004
	KDCoE	.581 ± .004	.680 ± .004	.628 ± .003	.730 ± .007	.837 ± .006	.778 ± .005	.482 ± .005	.515 ± .006	.499 ± .005	.611 ± .012	.653 ± .015	.632 ± .014
	BootEA	.507 ± .010	.718 ± .012	.603 ± .011	.660 ± .006	.850 ± .005	.745 ± .005	.389 ± .004	.561 ± .004	.474 ± .004	.640 ± .001	.806 ± .001	.716 ± .000
	GCNAlign	.338 ± .002	.589 ± .009	.451 ± .005	.414 ± .005	.698 ± .007	.542 ± .005	.230 ± .002	.412 ± .004	.319 ± .003	.257 ± .002	.455 ± .003	.351 ± .002
	AttrE	.481 ± .010	.671 ± .009	.569 ± .010	.535 ± .015	.746 ± .014	.631 ± .014	.403 ± .019	.572 ± .019	.483 ± .019	.466 ± .011	.644 ± .012	.549 ± .011
	IMUSE	.569 ± .006	.717 ± .010	.638 ± .008	.607 ± .013	.760 ± .014	.678 ± .013	.439 ± .002	.546 ± .004	.492 ± .003	.461 ± .003	.605 ± .005	.529 ± .004
	SEA	.280 ± .015	.530 ± .026	.397 ± .019	.360 ± .018	.651 ± .018	.494 ± .017	.225 ± .011	.399 ± .013	.314 ± .012	.297 ± .002	.500 ± .002	.395 ± .002
	RSN4EA	.393 ± .007	.595 ± .012	.487 ± .009	.579 ± .006	.759 ± .006	.662 ± .006	.293 ± .004	.452 ± .006	.371 ± .004	.495 ± .003	.672 ± .005	.578 ± .004
	MultiKE	.749 ± .004	.819 ± .005	.782 ± .004	.864 ± .007	.909 ± .005	.885 ± .006	.629 ± .002	.680 ± .002	.655 ± .002	.642 ± .003	.696 ± .003	.670 ± .003
	RDGCN	.755 ± .004	.854 ± .003	.800 ± .003	.847 ± .006	.919 ± .004	.880 ± .005	.640 ± .004	.732 ± .004	.683 ± .004	.715 ± .003	.787 ± .002	.748 ± .002

Stable matching (a.k.a. stable marriage)

The entity alignment between two KGs is stable when there does not exist another predicted aligned pair (e1; e2) of higher preference than those of e1 and e2 to their current matches. The preference can be calculated based on a similarity metric such as CSLS.

Table 6: Hits@1 w.r.t. distance metrics and alignment inference strategies on D-Y-15K (V1)

	Greedy	Greedy w/ CSLS	SM	SM w/ CSLS
MTransE	.463 ± .013	.550 ± .009	.694 ± .006	.697 ± .010
IPTransE	.313 ± .009	.339 ± .013	.370 ± .018	.369 ± .018
JAPE	.469 ± .009	.549 ± .009	.692 ± .015	.691 ± .015
KDCoE	.661 ± .013	.679 ± .000	.840 ± .024	.815 ± .031
BootEA	.739 ± .014	.741 ± .009	.783 ± .007	.782 ± .006
GCNAlign	.465 ± .012	.531 ± .008	.613 ± .008	.582 ± .010
AttrE	.668 ± .012	.778 ± .012	.845 ± .012	.857 ± .012
IMUSE	.392 ± .013	.448 ± .018	.520 ± .028	.518 ± .030
SEA	.500 ± .011	.557 ± .017	.647 ± .012	.650 ± .012
RSN4EA	.514 ± .003	.548 ± .003	.571 ± .002	.575 ± .004
MultiKE	.903 ± .004	.925 ± .003	.951 ± .003	.956 ± .002
RDGCN	.931 ± .004	.956 ± .002	.962 ± .002	.979 ± .001

In summary, existing approaches concentrate on developing more powerful embedding and interaction methods, but some methods for the alignment module can also improve performance.

Sect. 7 SUMMARY AND FUTUREDIRECTIONS

(i) RDGCN, BootEA and MultiKE achieve the most competitive performance. This suggests that incorporating both literal information and carefully-designed bootstrapping can help entity alignment.
(ii) For the embedding models designed for link prediction, we find that not all of them are suitable for entity alignment.
(iii) Currently, the alignment inference strategy receives little attention. Our preliminary results show that the CSLS distance metric and the stable matching strategy can bring performance improvement to all the approaches.
(iv) We also find that embedding-based and conventional entity alignment approaches are complementary to each other.
(v) For choosing appropriate approaches based on the available resources in real-world scenarios, Table 9 summarizes the required information of embedding-based and conventional entity alignment approaches in our experimental analysis.

Table 9: Summary of the required information of embedding-based and conventional entity alignment approaches

	MTransE	IPTransE	JAPE	KDCoE	BootEA	GCNAlign	AttrE	IMUSE	SEA	RSN4EA	MultiKE	RDGCN	LogMap	PARIS
Relation/attribute triples	•/	•/	•/○	○/○	•/	•/○	○/○	○/○	•/	•/	○/○	•/○	○/•	○/•
Pre-aligned ent./prop.	•/○	•/○	•/○	•/	•/	•/○	•/	•/	•/	•/	•/○	•/	○/○	○/○
Word embed./Google trans.				○/			○/				○/	○/	/△	/△

"•" means "mandatory", "○" means "optional", "△" means "mandatory for cross-lingual entity alignment", and blank means "not applicable".

Future Directions

Unsupervised entity alignment
Long-tail entity alignment
Large-scale entity alignment
Entity alignment in non-Euclidean spaces