CROSS-LINGUAL LONG-TAILED ENTITY ALIGNMENT IN KNOWLEDGE GRAPHS

Anonymous authors

Paper under double-blind review

ABSTRACT

Entity alignment (EA) models rely mostly on triples and structural information of Knowledge Graphs (KGs), but underperform on sparsely connected long-tailed entities. We address this gap by proposing a model, **ContrastEA**, that leverages pre-trained Language Models (LM), e.g. me5, to generate entity representations, followed by a novel contrastive learning approach that incorporates hard-negative mining strategies with *top-k* negatives per entity, alongside NT-Xent loss to separate challenging entity pairs. In addition, to address the under-representation of long-tailed entities in benchmark datasets, we curate a new dataset from DBpedia comprising long-tailed entities per language – Arabic, German, Portuguese, Italian, Hindi, Russian, and Japanese, each aligned to English (a total of 154,296 cross-lingual entity pairs). Our results demonstrate that ContrastEA outperforms the classic EA models on three benchmark datasets, improving Hits@1 by 6–20 percentage points, and achieves SOTA on the curated dataset over the long-tailed EA models.

1 Introduction

A Knowledge Graph (KG) is a structural network of relational facts stored in the form of triples \(\text{head entity, relation, tail entity} \) and serves as the backbone of various applications, including semantic search (Bast et al., 2016), question answering (Yasunaga et al., 2021), and recommender systems (Xian et al., 2019). Most KGs are created from single sources designed for specific purposes, representing niche domain knowledge, and therefore also with limited coverage. This limitation is further amplified in multilingual KGs (Kaffee et al., 2023), where the same entity may have more facts in one language than in another. The same entity may exist in different KGs in different forms. For example, dbr:Albert_Einstein in DBpedia (Lehmann et al., 2015) and wd:Q937 in Wikidata (Vrandečić & Krötzsch, 2014) refer to the same entity Albert Einstein. Therefore, integrating multiple KGs is the key to completeness and knowledge fusion for better inference and reasoning, as different KGs contain complementary knowledge.

Entity alignment (EA) is the core task of KG integration by finding entities across different KGs that refer to the same (real-world) entity. However, a principal challenge is the heterogeneity of the underlying schemas across different KGs. KG embedding based methods have proved to be an effective approach for EA (Chen et al., 2018; 2017; Pei et al., 2019a;b; Guo et al., 2019; Sun et al., 2020) and can be summarized into two categories: (i) the KG embedding model encodes the source and target KGs into two separate embedding spaces, and trains an alignment method with aligned initial seeds to learn the mapping between the same entities across KGs (Chen et al., 2018; 2017; Pei et al., 2019a); (ii) an alignment model is trained to embed the two KGs into one unified space using initial seed alignment (Li et al., 2018; Sun et al., 2017; 2018; Trisedya et al., 2019; Wang et al., 2018; Zhu et al., 2017). These models leverage the relational and structural information of the entities, assuming the KGs are strongly connected and equivalent entities have similar neighbourhood structures in different KGs. However, KGs are inherently sparse, with approximately 50% of the entities being *long-tailed*, having three connections or less to other entities (Guo et al., 2019). For instance, 54.7% of the entities in English DBpedia and 61.5% in the German DBpedia appear fewer than three times.

Although embedding-based EA models can potentially mitigate the challenge of heterogeneity, they inherently struggle with insufficient structural and neighbourhood information. This raises the open

 question How can we align long-tailed entities across KGs with fewer or no connections to their neighbouring entities? In this work, we propose a novel long-tailed EA model ContrastEA, that harnesses pretrained Language Models (LMs) to generate entity name representations, followed by a contrastive learning framework to align the same entities across different KGs. A hard-negative mining strategy is proposed to select the most informative top-k negatives per entity, ensuring the model to learn to differentiate between entities that are deceptively similar to the actual target entity. ContrastEA also includes the NT-Xent loss, which ensures the closeness of the positive pairs (i.e., same entities in different KGs) while pushing away the negative pairs in the embedding space.

Recent work (Guo et al., 2019) also uncovers that existing EA benchmark datasets fail to reflect the sparsity of the real-world KGs. To this end, the SRPRS benchmark (Guo et al., 2019) was introduced, with a more pronounced presence of long-tailed entities. However, entities with zero connections to other entities are not included in the SRPRS dataset. To address this and truly reflect the sparseness of KGs, we curate a multilingual EA benchmark dataset, LT-EA-25K, from DBpedia containing both long-tailed entities and entities with no links at all.

Prior works on long-tailed EA have used the latent representation of entity names derived from pretrained word embeddings as an additional signal to the structural information (Zeng et al., 2020), while classical EA models used entity name embeddings to initialize the feature matrices for structural representations (Wu et al., 2019a;b; Xu et al., 2019). In contrast, our proposed model **ContrastEA** uses only entity names, without depending on structural or relational information. We show that **ContrastEA** achieves state-of-the-art results on both the existing benchmarks as well as our curated dataset, ensuring its robustness for entities with varying connectivity. Our main contributions are:

- We propose a novel contrastive learning based EA approach ContrastEA that incorporates
 hard-negative mining strategies with top-k negatives per entity. It also exploits the NT-Xent
 loss to separate challenging entity pairs.
- We propose a multilingual long-tailed EA benchmark dataset **LT-EA-25K**, which comprises 25,000 long-tailed entities per language Arabic, German, Portuguese, Italian, Russian, and Japanese, each aligned to English (a total of 154,296 cross-lingual entity pairs) and 4,296 aligned pairs of Hindi–English ¹.
- Our detailed evaluation demonstrates that ContrastEA is robust for popular, long-tailed and entities with zero connectivity. Our results show that ContrastEA outperforms the classic EA models on three benchmark datasets, improving Hits@1 by 6–20 percentage points, as well as on the curated dataset by 10 percentage points on Hits@1 over the long-tailed EA models.

2 Methodology

Preliminaries Given a KG G=(E,R), where E and R are the set of entities and relations, respectively, the notation $\langle s,p,o\rangle\in T$ represents a triple belonging to the set of triples T in the KG, where $(s,o)\in E$ are the head and tail entities, and $p\in R$ represents the relation between them. Additionally, the entities in a KG are related to facts that are literals in the form of text, images, numbers, etc. Given a source KG $G_1=(E_1,R_1)$, a target KG $G_2=(E_2,R_2)$ and T_1 and T_2 are the sets of triples in the source and target KG, respectively. For training, the seed entity pairs are given by $S=\{(s,t)\mid s\in E_1,t\in E_2,s\leftrightarrow t\}$, where \leftrightarrow denotes equivalence, designating s and t as being the same entity (Zhao et al., 2020).

Problem Statement We consider the task of long-tailed EA as a supervised contrastive learning problem. Given a training set $\mathcal{D} = \{(s_i, t_i)\}_{i=1}^N$ of N aligned seed entities, where s_i and t_i are the i^{th} source and target entity, respectively. The goal is to learn a parametrised encoder $f_\theta: \mathcal{X} \to \mathbb{R}^d$ to generate a d-dimensional vector for the entity names of both s_i and t_i and subsequently train a contrastive learning model for EA. The training objective of the alignment model serves to encourage $f_\theta(s_i)$ to be similar to $f_\theta(t_i)$ and dissimilar to $f_\theta(t_i)$, for $i \neq j$.

¹The data is made available in the supplementary folder

2.1 ENTITY REPRESENTATIONS

As mentioned earlier, EA models that use structural, relational, or neighbourhood information of the entities are ill-suited for aligning long-tailed entities across KGs, as they frequently lack such information. Therefore, we focus on the most consistent and omnipresent feature: *entity names*. Entity descriptions and attributes are often unavailable for less connected entities, but names are informative and often sufficient to identify an entity (Yaghoobzadeh et al., 2018; Yaghoobzadeh & Schütze, 2017). We use a transformer encoder f_{θ} to map entity names to d-dimensional vectors. For an input text x, the encoder returns the hidden states $H \in \mathbb{R}^{m \times L \times d}$, where L is the length after padding/truncation, and m is the batch size. To obtain a single representation of the entity name, masked mean pooling is used where $M \in \{0,1\}^{m \times L}$ and the pooled embedding of the ith entity is:

$$e_i = \frac{\sum_j M_{i,j} H_{i,j}}{\sum_j M_{i,j}} \in \mathbb{R}^d$$

Leveraging entity embeddings for EA instead of simple string matching enables effective crosslingual entity alignment. These representations are crucial when the entity names differ across languages with different vocabularies, overcoming challenges of translation and transliteration.

2.2 Contrastive Learning

Our proposed **ContrastEA** model focuses on pulling together similar entities while pushing apart hard-to-distinguish entities, i.e., misleading similar entities, aiming for a more discriminative yet robustly aligned space. In prior work, during the training of the alignment module of the EA models, for each anchor entity e, the set of negative samples is sampled uniformly from the data without accounting for their informativeness with regard to the corresponding anchor entity. Studies show that informative negative samples can guide the training phase, as they are intuitively the ones that are mapped in close proximity, but instead should be far apart (Robinson et al., 2021; Chen et al., 2020). The proposed model consists of three variants of contrastive loss: (i) contrastive loss with top-k hard negatives, (ii) contrastive loss with aggregated top-k hard negatives, and (iii) In-batch NT-Xent loss, where all the negatives are used. Let $f_{\theta}(x) \in \mathbb{R}^d$ be the encoder output, i.e., the entity embeddings. Before computing similarities, the embeddings of the source and target entities are normalised: $z_i^s = \frac{f_{\theta}(s_i)}{\|f_{\theta}(s_i)\|}$, $z_i^t = \frac{f_{\theta}(t_i)}{\|f_{\theta}(t_i)\|}$.

Contrastive loss with top-k hard negatives The embedding of the source entity is considered as the query, and the set of target embeddings in the batch as keys. This loss selects top-k hardest negatives for each query (source entity) and employs a softmax over the positive and k negatives. Let the similarity matrix be,

$$A_{i,j} = z_i^{sT} \cdot z_j^t, \quad i \neq j$$

To ensure that the positive target is not selected as a negative, the diagonal is masked by setting its entries to $-\infty$. For each row i in $A_{i,j}$, we compute the set N_i of k indices with the largest value of $A_{i,j}$, where $j \neq i$. The top-k hardest negatives are selected $\mathcal{N}_i = \operatorname{TopK}(A_{i,j} \mid j \neq i)$, where $N_i = \{j_1, ..., j_k\}$. Then we compute the logits

$$L_{i} = \frac{1}{\tau} [A_{i,i}, A_{i,j_{1}},, A_{i,j_{k}}] \in \mathbb{R}^{(1+k)},$$

where au is the temperature. Next, the per-sample top-k cross-entropy loss is obtained as

$$\ell_{i,j}^{TopK} = -\log \frac{\exp(\text{sim}(z_i^{sT}, z_i^t)/\tau)}{\exp(\text{sim}(z_i^{sT}, z_i^t)/\tau) + \sum\limits_{j \in \mathcal{N}_i} \exp(\text{sim}(z_i^{sT}, z_j^t)/\tau)}$$

and the corresponding batch loss is $\frac{1}{m} \sum_{i,j=1}^{m} \ell_{i,j}, i \neq j$, where m is the batch size.

Contrastive loss with aggregated top-k hard negatives In this loss function, for each source entity (query), after selecting the relevant top-k hard negative samples as described above, we compute their mean similarities to obtain a single aggregated negative score per entity. This is mathematically given by

$$\mathrm{neg}_i = \frac{1}{k} \sum_{j \in \mathcal{N}_i} A_{i,j}$$

where neg_i is the aggregated negative score for source entity i, k is the number of negatives, and \mathcal{N}_i is the list of top-k hardest negatives for the entity i.

NT-Xent loss with in-batch negatives Our contrastive learning approach also explores a diverse sample of negatives for a given anchor entity using the NT-Xent loss with in-batch negatives (Chen et al., 2020). The goal of this loss is to maximise the similarity between positive pairs (true source entity – target entity pair) and minimise the similarity between negative pairs (incorrect source entity – target entity pair). Given a batch of m normalised pairs of source and target entities $\{z_i^s, z_i^t\}_{i=1}^m$, a similarity matrix $A_{i,j}$ of logits is generated, where $A \in \mathbb{R}^{m \times m}$ with entries $A_{i,j} = \frac{z_i^{sT}, z_j^t}{\tau}$, $i \neq j$, where τ is the temperature. Here, all pairs of cosine similarity are scaled by a temperature value before computing the cross-entropy loss. For each query i (source entity), the positive key (target entity) is j = i, and for each source entity, the cross-entropy loss is given by

$$\ell_{i,j}^{\text{NT-Xent}} = -\log \frac{\exp(A_{i,i})}{\sum_{j=1}^{m} \exp(A_{i,j})} = -A_{i,i} + \log \left(\sum_{j=1}^{m} A_{i,j}\right),$$

where m is the batch size, and the batch loss is given by $\frac{1}{m} \sum_{i,j=1}^{m} \ell_{i,j}, i \neq j$.

Here, the embedding of the source entity is considered for each input, and we then compute the cosine similarity between every pair of target entity vectors. Since each entity in the source KG is similar to exactly one other entity in the target KG (positive pair), it can be considered as a classification task where similarity scores correspond to the probability distribution, in which the correct targets approach 1.0 and the rest will be close to 0.0.

Why Hard Negatives Matter? For the NT-Xent loss, the partial derivative of $\ell_i^{\text{NT-Xent}}$ with respect to $\sin(z_i^s, z_{n-}^t)$ for a specific negative n^- is proportional to a softmax weight

$$\frac{\partial \ell_{i,j}^{\text{NT-Xent}}}{\partial \text{sim}(z_i^s, z_{n^-}^t)} \; \alpha \frac{\exp(\text{sim}(z_i^s, z_{n^-}^t)/\tau)}{\sum_{i \neq a} \exp(\text{sim}(z_i^s, z_{a^-}^t)/\tau)}$$

Therefore, negatives with larger similarity (hard negatives) will have larger gradients; however, the collective impact of many easy negatives might overpower the impact of a few hard negatives in the training. Therefore, restricting the denominator to only the hard negatives \mathcal{N}_i , as demonstrated in $\ell_{i,j}^{\text{TopK}}$, will explicitly increase the influence of the negatives with large similarity by concentrating the gradients.

3 EXPERIMENTS AND RESULTS

3.1 Datasets

Benchmark Datasets The most commonly used cross-lingual EA benchmark datasets are DBP15K (Sun et al., 2017), SRPRS (Guo et al., 2019), and DBP5L (Chakrabarti et al., 2022) (detailed statistics are provided in Table 1). Although they are sampled from the real-world KGs DBpedia, Wikidata and YAGO, their entity distributions do not reflect the actual distributions in the original KGs. These real-world KGs typically follow a power-law degree distribution with more than 50% of the entities having degrees less than 3. In contrast, the DBP15K dataset, which contains aligned entities from three language pairs ZH–EN, JA–EN, and FR–EN, within which 21% to 43% of the entities have a degree less than or equal to 3. Similarly, in DBP5L, 33–56% of entities have a degree less than or equal to 3. This over-representation of the connectivity of the entities in the DBP15k dataset is rectified in SRPRS dataset, in which only 77–85% of entities have degrees less than or equal to 3. The detailed degree distribution of the entities in these benchmarks datasets is shown in Figure 3 and in Table 7 in Appendix A.1.

LT-EA-25k dataset Although the SRPRS dataset includes low-degree entities, it is restricted to only two languages – German and French – and their corresponding alignment to English, limiting its usage in the evaluation of extensive cross-lingual EA. Moreover, the connectivity and completeness of the KGs vary from language to language. For instance, English or German DBpedia is denser

Table 1: Statistics of the benchmark datasets DBP15k, SRPRS, and DBP5L

Dataset	Lang.	Entities	Relations	Rel. Triples	Attr. Triples
DBP15K	ZH	19,388	1,701	70,414	379,684
(ZH-EN)	EN	19,572	1,323	95,142	567,755
DBP15K	JA	19,814	1,299	77,241	354,619
(JA-EN)	EN	19,780	1,153	93,484	497,230
DBP15K	FR	19,661	903	105,998	528,665
(FR-EN)	EN	19,993	1,208	115,722	576,543
SRPRS	EN	15,000	222	38,363	62,715
(EN-DE)	DE	15,000	120	37,377	142,506
SRPRS	EN	15,000	221	36,508	70,750
(EN-FR)	FR	15,000	177	33,532	56,344
	EN	13,996	831	80,167	_
	FR	13,176	178	49,015	_
DBP-5L	ES	12,382	144	54,066	_
	JA	11,805	128	28,774	_
	EL	5,231	111	13,839	_

compared to Hindi or Arabic DBpedia. Therefore, a dataset with more languages would also change the proportion of the degree distribution of the entities. Also, none of these benchmark datasets consider *dangling entities* in the KGs, i.e., entities with no connections to other entities.

To address these limitations, we introduce a new cross-lingual long-tailed entity alignment dataset LT-EA-25k comprising seven languages – Arabic, German, Hindi, Italian, Japanese, Portuguese, and Russian and their corresponding alignment with English curated from DBpedia. The proportion of entities with a degree ≤ 3 varies between 42% and 81% depending on the language, reflecting the sparsity and heterogeneity of multilingual KGs. To curate the proposed dataset, we follow these steps for each language-specific KG:

- We perform degree profiling for all the entities i.e., we compute the outdegree, indegree, and the total degree of each entity.
- We divide the entities into 6 buckets based on their degree values, i.e. $\{1, 2, 3, 4, 5, >5\}$.
- Next, we select entities from each bucket using a stratified random sampling to ensure a true representative distribution of the entities in each language edition of DBpedia, and thereby avoiding high connectivity bias as observed in the DBP15k and DBP5L datasets.
- The corresponding aligned entities from English DBpedia are extracted by exploiting the owl:sameAs links using SPARQL queries.

We also include smaller language editions of DBpedia, namely Hindi. Therefore, LT-EA-25k contains only 4,296 aligned entities between Hindi and English while maintaining the degree distribution of the entities. The Arabic edition of DBpedia is also very sparse, as evidenced in our dataset, which contains 11,617 entities with degree 1. We also incorporated ~1,400 dangling entities from the Arabic DBpedia. A detailed statistics of our dataset is provided in Table 2 and the degree distribution in Figure 1 and in Table 8 of Appendix A.1. Unlike prior works, we also include medium and low-resource language KGs in our dataset, instead of merely high-resource language-specific KGs that are well-connected. Overall, we provide a robust dataset mirroring the real-world KGs by considering multilinguality, heterogeneity, and sparsity.

3.2 EXPERIMENTAL SETUP

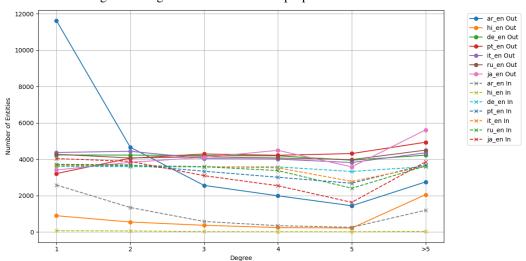
Implementation In our experiments, we use the mE5-base 2 encoder (Wang et al., 2024) to produce normalised text embeddings for entity names. The model is initialised with XLM-RoBERTa-base model, has 12 layers, and the embedding dimensionality is 768. It supports ~ 100 languages, which is crucial for cross lingual EA. For robust training of **ContrastEA**, we perform a hyperparameter search for the following values. For the learning rate: $\{0.00001, 0.00002, 0.00003, 0.00004, 0.00004, 0.00002, 0.00003, 0.00004, 0.00002, 0.0$

https://huggingface.co/intfloat/multilingual-e5-base

Table 2: Statistics of proposed long-tailed EA dataset, LT-EA-25k

Lang. Pairs	Aligned Entities	Triples Source Lang.	Triples Target Lang.	Attr. Source Lang.	Attr. Target Lang.
AR-EN	25,000	66,613	166,740	29,293	91,672
DE-EN	25,000	107,418	117,647	146,926	94,519
HI–EN	4,296	30,921	31,425	4,960	8,600
IT-EN	25,000	106,589	126,566	87,206	83,034
JA-EN	25,000	115,373	130,072	56,408	69,939
PT-EN	25,000	111,742	111,902	75,660	101,007
RU-EN	25,000	101,047	132,867	60,343	85,219

Figure 1: Degree Distribution of the proposed LT-EA-25k dataset



 $0.00005,\ 0.0001,\ 0.0002,\ 0.0003,\ 0.0004,\ 0.0005,\ 0.001,\ 0.002,\ 0.003,\ 0.004,\ 0.005\}$, batch size: $\{16,\ 32,\ 64\}$, epochs: $\{3,\ 5,\ 10,\ 15,\ 20\}$, warmup ratio: $\{0.1,\ 0.5,\ 0.8\}$, top-k (for hard negatives in contrastive learning): $\{3,\ 5,\ 10,\ 15\}$, weight decay: $\{0.1,\ 0.01,\ 0.001,\ 0.0001\}$, and temperature: $\{0.1,\ 0.01,\ 0.001,\ 0.0001\}$. The selected settings are: learning rate: 2e-5, batch size: 32, epochs: 5, warmup ratio: 0.1, top-k: 5, weight decay: 0.01, and temperature: 0.1. We also fix the random seed at 42 for reproducibility. Topk = 5 indicates that during training for each source entity, the most similar 5 target entities are selected as negative samples. We use 30% of the aligned entities as the training set and the remaining 70% as the test set. Following the convention, we report results for Hits@1 and Hits@10 for all three benchmark datasets – DBP15k, SRPRS, and DBP5L, as well as our curated dataset LT-EA-25k.

3.3 RESULTS

As shown in Tables 3 and 4, our proposed model ContrastEA achieves SOTA results in both SR-PRS and DBP15k datasets for both Hits@1 (exact match) and Hits@10. On the SRPRS dataset, for Hits@1 ContrastEA achieves a steep improvement of 20 percentage points for EN-FR language pair, and ~10 percentage points for EN-DE over the previous best method, DAT. The DAT Zeng et al. (2020) model focuses on the alignment of long-tailed entities and uses entity names as a feature, together with the neighbourhood information of entities. Since Hits@1 is equivalent to precision, this result highlights that most of the entities are correctly aligned at the highest rank by the ContrastEA model. On the other hand, Hits@10 reaches 0.978 (EN-FR) and 0.99 (EN-DE), indicating that the majority of the correct alignments are captured within the top 10 candidates, demonstrating robust recall. Although all baseline models use the structural and relational information of the entities, our empirical results demonstrate that the entity representations obtained from the pre-trained LM coupled with the contrastive objectives are more effective. Similarly, in the DBP15k dataset,

324	0	0	Л
	J	4	+

327

328

330

331

332

333

334

335

GCN

JAPE

RDGCN

GM-Align

ContrastEA

HGCN

DAT

Table 3: Results on SRPRS dataset									
Methods	EN	-FR		-DE					
Methous	Hits@1	Hits@10	Hits@1	Hits@10					
MTransE	0.251	0.551	0.312	0.586					
IPTransE	0.255	0.557	0.313	0.592					
BootEA	0.313	0.629	0.442	0.701					
RSNs	0.348	0.637	0.497	0.733					
MuGNN	0.131	0.342	0.245	0.431					
KECG	0.298	0.616	0.444	0.707					
TransEdge	0.4	0.675	0.556	0.753					

0.345

0.562

0.769

0.77

0.899

0.978

0.155

0.256

0.675

0.67

0.627

0.758

0.96

0.253

0.32

0.783

0.763

0.677

0.876

0.97

0 464

0.599

0.884

0.86

0.95

0.99

Table	4:	Results	on	DBP-1	5k	data	set.
hode		7H FN		TA	EN		

Methods	ZH	-EN	JA	-EN	FR-EN		
	H@1	H@10	H@1	H@10	H@1	H@10	
MTransE	0.308	0.614	0.279	0.575	0.244	0.556	
IPTransE	0.406	0.735	0.367	0.693	0.333	0.685	
JAPE	0.412	0.745	0.363	0.685	0.324	0.667	
AlignE	0.472	0.792	0.448	0.789	0.481	0.824	
GCN-Align	0.413	0.744	0.399	0.745	0.373	0.745	
SEA	0.424	0.796	0.385	0.783	0.400	0.797	
RSN	0.508	0.745	0.507	0.737	0.516	0.768	
MuGCN	0.494	0.844	0.501	0.857	0.495	0.870	
GCN	0.487	0.790	0.507	0.805	0.508	0.808	
GAT	0.418	0.667	0.446	0.695	0.442	0.731	
R-GCN	0.463	0.734	0.471	0.754	0.469	0.758	
AliNet	0.539	0.826	0.549	0.831	0.552	0.852	
ContrastEA	0.602	0.778	0.774	0.879	0.96	0.992	

336 337

338

339 340

347 348 349

350 351 352

368

369

370

371 372

373

374

375

376 377

Table 5: Results on DBP5L dataset. LangPair ContrastEA AlignKGC RAGA RNM RDGCN EL-EN 0.838 0.758 0.749 0.713 0.89**EL-ES** 0.833 0.798 0.794 0.747 0.88 EL-FR 0.821 0.691 0.724 0.727 0.85 0.644 EL-JA 0.748 0.644 0.683 0.79 JA-EN 0.765 0.591 0.645 0.582 0.856 JA-ES 0.743 0.583 0.65 0.6 0.827 JA-FR 0.739 0.645 0.706 0.602 0.813 ES-FR 0.895 0.809 0.849 0.871 0.939 **ES-EN** 0.933 0.857 0.88 0.878 0.977 0.945 0.905 **EN-FR** 0.77 0.812 0.832

the large improvement in the FR-EN language pair shows that our proposed model is better suited for cross-lingual alignment of closely related languages, confirming the effectiveness of the entity embeddings. For both JA-EN and ZH-EN, ContrastEA outperforms the second best model AliNet, by 22.5 and 6 percentage points, respectively, indicating that the model can overcome linguistic barriers such as different scripts. For Hits@10, ContrastEA achieves comparable results for ZH-EN language pair while outperforming for JA-EN and FR-EN. The GCN based models achieve high performance for Hits@10 but not for Hits@1, indicating the structural information of the entities is alone insufficient to distinguish between similar deceptive entities, while contrastive learning ensures language separation. Table 5 shows that ContrastEA also achieves the SOTA for all the language pairs in DBP5L dataset. While RDGCN (Wu et al., 2019a) leverages the relational information, AlignKGC (Chakrabarti et al., 2022), RAGA (Zhu et al., 2021a), and RNM (Zhu et al., 2021b) includes textual information. This dataset consists of several European languages, and the results confirm that ContrastEA is capable of capturing the fine-grained semantic characteristics of closely-related languages. The model also achieves the best results for the distant language pairs, namely, JA-FR, JA-ES and JA-EN. These language pairs are most challenging due to their linguistic differences as well as the scripts. ContrastEA shows an improvement of 5-9 percentage points over the existing best baseline model, AlignKGC. Therefore, it can be concluded that ContrastEA is robust across different language pairs with linguistic variations. Intuitively, it can be presumed that the KG embedding based or GNN-based EA models rely mostly on the structural information, precisely the local neighbourhood, and therefore fail to capture the true representations when entities are sparse or noisy. Furthermore, the topk hard negative contrastive learning plays a vital role in separating the true pairs from the deceptive ones. Further results are available in Appendix A.2.

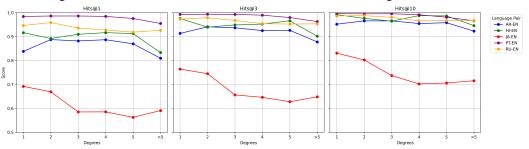
The results in Table 6 highlight a clear performance gap between ContrastEA and DAT across almost all language pairs. DAT has relatively very low, with Hits@ $1 \le 1.44$ for all language pairs, suggesting that it has trouble successfully aligning entities on this dataset. We trained our dataset on the DAT ³ model as it is focused on long-tailed entities, and also, as shown in Table 3, it achieves the second-best performance after ContrastEA. As mentioned in Section 3.1, the proposed LT-EA-25k

³https://github.com/DexterZeng/DAT

Table 6: Results for the LT-EA-25k dataset

Model/		ContrastE	A	DAT		
Language Pair	Hits@1	Hits@3	Hits@10	Hits@1	Hits@10	
AR-EN	81.42	88.39	92.67	1.44	4.78	
HI-EN	86.17	92.82	95.55	1.36	6.56	
IT-EN	95.72	96.91	97.67	0.75	3.93	
RU-EN	90.12	94.10	96.09	0.56	2.68	
JA-EN	53.33	59.73	65.95	0.97	3.40	
PT-EN	96.11	97.84	98.38	1.05	4.60	
DE-EN	97.31	98.49	99.05	0.59	3.39	

Figure 2: Results of ContrastEA on LT-EA-25k dataset w.r.t. degree distribution



dataset includes a large proportion of long-tailed entities with connectivity less than 3. Since the existing models, including DAT, rely heavily on the neighbourhood connection, the models are incapable of handling sparse KG; therefore do not produce entity embeddings that are well represented. Moreover, the language pair AR-EN is particularly challenging as it contains 1400 dangling entities, for which there is no neighbourhood information. Despite this, ContrastEA achieves a high Hits@1 (81.42%), whereas DAT achieves a low Hits@1. The hardest language pair for ContrastEA appears to be JA-EN, which is due to the linguistic differences, besides a larger number of long-tailed entities. Similar observation has also been made in general for JA-EN, also in DBP15K datasets, as illustrated in Table 4. However, the alignment of the European languages IT, PT, DE, and RU with English achieves stronger results, indicating higher language similarities and effective separation of dissimilar entities from the similar ones. Another advantage of ContrastEA is that no training is required to generate the entity embeddings.

We report the ablation studies of ContrastEA on the degree distribution of the entities in the LT-EA-25k dataset, as shown in Figure 2. We reported results on Hits@1, Hits@3 and Hits@10 across degrees (1, 2, 3, 4, 5, >5). The model achieves remarkable performance on Hits@1 with Degree = 1,2, with HI and PT achieving the highest values 0.9165 (for Deg=1), and 0.8924 (for Deg=2) and 0.9844 (for Deg=1), 0.9863 (for Deg=2), respectively. These values show that the model is robust with sparsely connected entities and also agnostic towards their linguistic differences. We observe that with Hits@1 and Hits@3, the performance of the model drops a little across most of the languages, with entities having degrees more than 5. However, this is an aggregated performance of the model for all entities. Overall, high values of Hits@1 and Hits@10 across different degrees indicate that ContrastEA is consistent with high precision and recall. Since a major proportion of the curated LT-EA-25k dataset is sparse, replicating the entity distribution of real-world KGs, we infer from the ablation studies that ContrastEA is capable of learning ambiguous (learning deceptive entities in TopK loss) entity patterns in sparse settings. Also, since no training is required to generate the entity representations, and it also does not consider the neighbourhood or the relational information of the entities, the model is scalable to larger multilingual KGs comprising long-tailed entities and low-resource languages, as well as with popular entities and high-resource languages.

4 RELATED WORK

Recent research on entity alignment (EA) broadly falls into two categories: semantic matching-based models and graph neural network (GNN)-based models (Zeng et al., 2021; Sun et al.).

Semantic matching-based models aim to embed entities in a low-dimensional space such that semantically similar entities are close together. Early approaches like MTransE (Chen et al., 2017), IPTransE (Zhu et al., 2017), rely on translational embeddings (e.g., TransE) and linear transformations for alignment. Models such as JAPE (Sun et al., 2017) incorporate attributes or textual descriptions, semi-supervised or iterative frameworks like BootEA (Sun et al., 2018), to improve alignment accuracy. Recent methods, e.g., BERT-INT (Tang et al., 2021), leverage pre-trained language models to encode entity names and descriptions.

GNN-based models exploit the graph structure of knowledge graphs, encoding nodes based on neighbour information to capture local and global structural patterns. GCN-Align (Xu et al., 2019) pioneered this direction, followed by models like RDGCN (Wu et al., 2019a), which include relational information. HMAN (Yang et al., 2019) further integrates entity attributes, textual descriptions, and pre-trained embeddings within GNN frameworks. Other models include AttrGNN (Liu et al., 2020), and REA (Pei et al., 2020), which collectively improve alignment by leveraging graph topology and semantics. Recent works (Zeng et al., 2020) have also used entity names as features together with structural and neighbourhood information.

5 CONCLUSION

In this paper, we introduce a novel cross-lingual entity alignment method, ContrastEA, which achieves state-of-the-art across the three widely used benchmark datasets and our proposed LM-EA25k dataset. The experimental results demonstrate consistent and robust performance across entities with varying degrees - popular and under-represented entities in a KG, namely long-tailed, and dangling entities. By leveraging the multilingual LM, the model achieves high precision (Hits@1) across diverse language pairs with linguistic differences, including different scripts, word order, and typologically distant languages from different language families. Therefore, it can be concluded that our model does not overfit to the language-specific structures and thereby improves the generalizability of our approach, advancing the knowledge fusion on a broader spectrum. Our proposed dataset would also pave the way for future research in long-tailed and dangling entity alignment across different KGs.

In the future, we intend to focus on incorporating a short description of the entities in the model, which would help us disambiguate between highly correlated yet deceptive entities. As KGs are inherently sparse and are continuously evolving, our focus would be on multilingual KG completion with the help of cross-lingual EA.

REFERENCES

Hannah Bast, Björn Buchhold, and Elmar Haussmann. Semantic search on text and knowledge bases. *Found. Trends Inf. Retr.*, 10(2–3):119–271, June 2016. ISSN 1554-0669. doi: 10.1561/1500000032. URL https://doi.org/10.1561/1500000032.

Soumen Chakrabarti, Harkanwar Singh, Shubham Lohiya, Prachi Jain, et al. Joint completion and alignment of multilingual knowledge graphs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11922–11938, 2022.

Muhao Chen, Yingtao Tian, Mohan Yang, and Carlo Zaniolo. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2017.

Muhao Chen, Yingtao Tian, Kai-Wei Chang, Steven Skiena, and Carlo Zaniolo. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, pp. 3998–4004, 2018.

- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.
 - Lingbing Guo, Zequn Sun, and Wei Hu. Learning to exploit long-term relational dependencies in knowledge graphs. In *International conference on machine learning*, pp. 2505–2514. PMLR, 2019.
 - Lucie-Aimée Kaffee, Russa Biswas, C. Maria Keet, Edlira Kalemi Vakaj, and Gerard de Melo. Multilingual Knowledge Graphs and Low-Resource Languages: A Review. *Transactions on Graph Data and Knowledge*, 1(1):10:1-10:19, 2023. doi: 10.4230/TGDK.1.1.10. URL https://drops.dagstuhl.de/entities/document/10.4230/TGDK.1.1.10.
 - Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. Dbpedia a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015. URL http://dblp.uni-trier.de/db/journals/semweb/semweb6.html#LehmannIJJKMHMK15.
 - Shengnan Li, Xin Li, Rui Ye, Mingzhong Wang, Haiping Su, and Yingzi Ou. Non-translational alignment for multi-relational networks. In *IJCAI*, pp. 4180–4186, 2018.
 - Zhiyuan Liu, Yixin Cao, Liangming Pan, Juanzi Li, and Tat-Seng Chua. Exploring and evaluating attributes, values, and structures for entity alignment. *arXiv preprint arXiv:2010.03249*, 2020.
 - Shichao Pei, Lu Yu, Robert Hoehndorf, and Xiangliang Zhang. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference. In *The world wide web conference*, pp. 3130–3136, 2019a.
 - Shichao Pei, Lu Yu, and Xiangliang Zhang. Improving cross-lingual entity alignment via optimal transport. International Joint Conferences on Artificial Intelligence, 2019b.
 - Shichao Pei, Lu Yu, Guoxian Yu, and Xiangliang Zhang. Rea: Robust cross-lingual entity alignment between knowledge graphs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2175–2184, 2020.
 - Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021.
 - Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proceedings of the VLDB Endowment*, 13(11).
 - Zequn Sun, Wei Hu, and Chengkai Li. Cross-lingual entity alignment via joint attribute-preserving embedding. In *International semantic web conference*, pp. 628–644. Springer, 2017.
 - Zequn Sun, Wei Hu, Qingheng Zhang, and Yuzhong Qu. Bootstrapping entity alignment with knowledge graph embedding. In *Ijcai*, volume 18, 2018.
 - Zequn Sun, Chengming Wang, Wei Hu, Muhao Chen, Jian Dai, Wei Zhang, and Yuzhong Qu. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 222–229, 2020.
 - Xiaobin Tang, Jing Zhang, Bo Chen, Yang Yang, Hong Chen, and Cuiping Li. Bert-int: a bert-based interaction model for knowledge graph alignment. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pp. 3174–3180, 2021.
 - Bayu Distiawan Trisedya, Jianzhong Qi, and Rui Zhang. Entity alignment between knowledge graphs using attribute embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 297–304, 2019.

- Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September 2014. ISSN 0001-0782. doi: 10.1145/2629489. URL https://doi.org/10.1145/2629489.
 - Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*, 2024.
 - Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 349–357, 2018.
 - Y Wu, X Liu, Y Feng, Z Wang, R Yan, and D Zhao. 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, 2019a.
 - Y Wu, X Liu, Y Feng, Z Wang, and D Zhao. Jointly learning entity and relation representations for entity alignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 240–249. Association for Computational Linguistics, 2019b.
 - Yikun Xian, Zuohui Fu, S. Muthukrishnan, Gerard de Melo, and Yongfeng Zhang. Reinforcement knowledge graph reasoning for explainable recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SI-GIR'19, pp. 285–294, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361729. doi: 10.1145/3331184.3331203. URL https://doi.org/10.1145/3331184.3331203.
 - Kun Xu, Liwei Wang, Mo Yu, Yansong Feng, Yan Song, Zhiguo Wang, and Dong Yu. Crosslingual knowledge graph alignment via graph matching neural network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3156–3161, 2019.
 - Yadollah Yaghoobzadeh and Hinrich Schütze. Multi-level representations for fine-grained typing of knowledge base entities. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers,* 2017.
 - Yadollah Yaghoobzadeh, Heike Adel, and Hinrich Schütze. Corpus-level fine-grained entity typing. J. Artif. Intell. Res., 61:835–862, 2018.
 - Hsiu-Wei Yang, Yanyan Zou, Peng Shi, Wei Lu, Jimmy Lin, and Xu Sun. Aligning cross-lingual entities with multi-aspect information. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4431–4441, 2019.
 - Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 535–546, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.45. URL https://aclanthology.org/2021.naacl-main.45/.
 - Kaisheng Zeng, Chengjiang Li, Lei Hou, Juanzi Li, and Ling Feng. A comprehensive survey of entity alignment for knowledge graphs. *AI Open*, 2:1–13, 2021.
 - Weixin Zeng, Xiang Zhao, Wei Wang, Jiuyang Tang, and Zhen Tan. Degree-aware alignment for entities in tail. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 811–820, 2020.
 - Xiang Zhao, Weixin Zeng, Jiuyang Tang, Wei Wang, and Fabian M Suchanek. An experimental study of state-of-the-art entity alignment approaches. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2610–2625, 2020.

Hao Zhu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. Iterative entity alignment via joint knowledge embeddings. In *IJCAI*, volume 17, pp. 4258–4264, 2017.

Renbo Zhu, Meng Ma, and Ping Wang. Raga: relation-aware graph attention networks for global entity alignment. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 501–513. Springer, 2021a.

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

A APPENDIX

A.1 DEGREE DETAILS

Table 7: Node outdegree distribution across different datasets and languages.

Dataset / Lang		Outdeg=2	Outdeg=3			Outdeg>5				
			SRPRS							
DBP	3834	5527	2898	1172	646	923				
WD	5155	5319	1765	784	493	1484				
SRPRS										
DBP	6679	4005	1816	988	618	894				
YG	6687	4000	1480	832	536	1465				
			SRPRS							
EN	5101	4285	2227	1384	818	1185				
DE	6675	3670	1534	1063	636	1422				
			SRPRS							
EN	5816	4464	1804	1102	696	1118				
FR	5889	5229	1702	839	498	843				
DBP-15K										
FR	1126	1265	1105	1221	1359	8924				
EN	967	1062	1053	1102	1228	9588				
			DBP-15K							
JA	1385	1618	2016	2173	2457	5351				
EN	1059	1363	1493	1718	2233	7134				
			DBP-15K							
ZH	1704	2227	2499	2334	1776	4460				
EN	954	1501	1742	1805	1760	7238				
DBP5L Dataset										
EL	836	695	450	376	350	830				
EN	1097	1465	1524	1322	1175	5691				
ES	1165	1216	889	867	695	4046				
FR	841	1236	994	720	670	3834				
JA	1181	871	604	510	374	2152				

A.2 RESULTS

Figure 3: Degree distribution of all the benchmark datasets DBP15k (source-target), DBP5L (source-target), SRPRS (source-target)

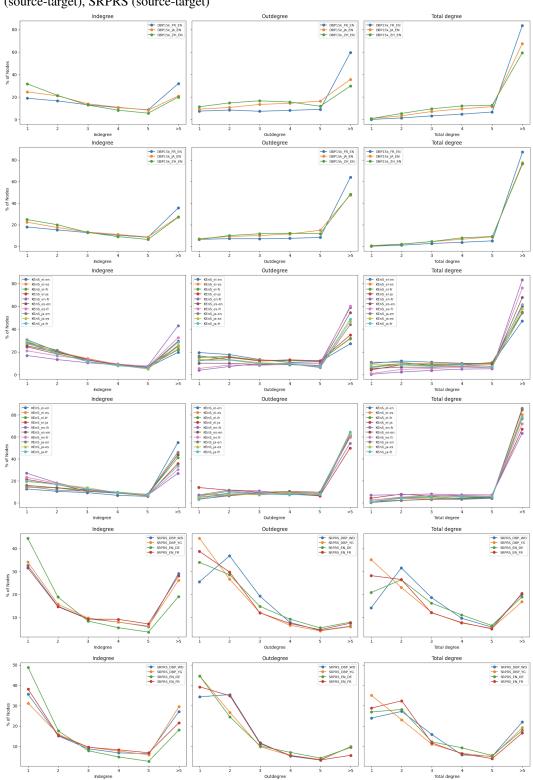


Table 8: Degree Distribution of the proposed LT-EA-25k dataset

Lang Pairs			Outdegree	Distribution	n				Indegree I	Distribution		
Langrans	Deg=1	Deg=2	Deg=3	Deg=4	Deg=5	Deg>5	Deg=1	Deg=2	Deg=3	Deg=4	Deg=5	Deg>5
ar_en	11,617	4,654	2,557	1,986	1,438	2,748	2,578	1,341	574	341	252	1,192
hi_en	890	544	363	239	212	2,048	68	55	15	10	8	31
de_en	4,231	4,224	4,197	4,185	3,943	4,220	3,601	3,585	3,577	3,575	3,323	3,590
pt_en	3,210	4,052	4,285	4,214	4,308	4,931	3,720	3,683	3,336	3,005	2,677	3,698
it_en	4,368	4,433	4,034	4,001	3,816	4,348	3,647	3,644	3,585	3,528	2,776	3,607
ru_en	4,267	4,067	4,131	4,065	3,973	4,497	3,708	3,652	3,585	3,372	2,401	3,623
ja_en	3,405	3,826	4,098	4,485	3,576	5,610	4,033	3,879	3,094	2,539	1,626	3,850

Table 9: ContrastEA results across different benchmarking datasets and language pairs.

Dataset / Pair	Hits@1	Hits@3	Hits@10	Accuracy							
	DBP5L Dataset										
el_en	0.8865	0.9451	0.9682	0.8865							
el_es	0.8844	0.9396	0.9570	0.8844							
el_fr	0.8508	0.9043	0.9416	0.8508							
el_ja	0.7880	0.8520	0.8962	0.7880							
en_fr	0.9459	0.9663	0.9742	0.9459							
es_en	0.9778	0.9908	0.9956	0.9778							
es_fr	0.9396	0.9618	0.9737	0.9396							
ja_en	0.8569	0.9154	0.9500	0.8569							
ja_es	0.8278	0.8907	0.9272	0.8278							
ja_fr	0.8134	0.8707	0.9097	0.8134							
	SRF	PRS Datase	et								
en_de	0.9777	0.9869	0.9909	0.9777							
en_fr	0.9612	0.9729	0.9786	0.9612							
DBP-15K Dataset											
fr₋en	0.9607	0.9842	0.9921	0.9607							
ja_en	0.7748	0.8398	0.8796	0.7748							
zh_en	0.6017	0.6962	0.7786	0.6017							

Table 10: Results on DBP15k dataset with mE5 + MUSE.

method	src_lang	tgt_lang	train_size	Hits@1	Hits@5	Hits@10	Hits@1	Hits@5	Hits@10
unsupervised	fr	en	30	95.183	98.502	98.941	97.243	99.370	99.609
unsupervised	fr	en	50	94.594	98.371	98.812	96.916	99.292	99.559
unsupervised	fr	en	70	93.486	98.155	98.644	96.243	99.289	99.511
unsupervised	fr	en	90	91.933	98.067	98.467	95.000	99.333	99.600
unsupervised	ja	en	30	70.309	81.840	84.564	79.365	87.960	90.071
unsupervised	ja	en	50	67.997	80.349	83.159	77.944	87.016	89.194
unsupervised	ja	en	70	65.194	78.583	81.431	75.847	85.580	87.755
unsupervised	ja	en	90	58.883	73.284	76.447	70.390	82.100	84.455
unsupervised	zh	en	30	51.129	67.151	72.020	64.999	78.592	82.456
unsupervised	zh	en	50	48.265	65.054	70.159	63.071	77.288	81.469
unsupervised	zh	en	70	45.767	63.301	68.684	61.269	76.368	80.612
unsupervised	zh	en	90	45.839	62.886	67.718	61.745	75.168	80.537
supervised	fr	en	30	94.868	98.388	98.884	97.310	99.351	99.599
supervised	fr	en	50	94.180	98.211	98.745	97.010	99.279	99.573
supervised	fr	en	70	92.752	97.977	98.577	96.398	99.266	99.533
supervised	fr	en	90	90.600	97.800	98.467	95.133	99.267	99.533
supervised	ja	en	30	68.803	80.190	83.193	79.672	87.903	90.061
supervised	ja	en	50	65.067	77.944	81.196	77.769	86.788	89.073
supervised	ja	en	70	59.206	74.322	78.157	75.196	85.064	87.464
supervised	ja	en	90	51.884	67.362	71.669	68.977	80.686	83.917
supervised	zh	en	30	49.713	64.980	70.107	65.850	78.295	82.265
supervised	zh	en	50	44.955	61.624	67.279	63.701	77.087	81.187
supervised	zh	en	70	38.687	57.405	63.815	61.068	75.854	80.121
supervised	zh	en	90	37.852	54.631	60.805	60.067	74.698	78.993

Table 11: Results on DBP15k dataset on experiments that uses me5 to generate the embeddings followed by MUSE alignment

method	src_lang	tgt_lang	Hits@1	Hits@10	Hits@1	Hits@10
unsupervised	fr	en	95.183	98.941	97.243	99.609
unsupervised	ja	en	70.309	84.564	79.365	90.071
unsupervised	zh	en	51.129	72.020	64.999	82.456
supervised	fr	en	94.868	98.884	97.310	99.599
supervised	ja	en	68.803	83.193	79.672	90.061
supervised	zh	en	49.713	70.107	65.850	82.265