HLMEA: Unsupervised Entity Alignment Based on Hybrid Language Models

Xiongnan Jin[1], Zhilin Wang[2], Jinpeng Chen[3,4], Liu Yang[5], Byungkook Oh[6], Seung-won Hwang[7], Jianqiang Li[1]*

[1]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University
[2]Alibaba Group
[3]School of Computer Science, Beijing University of Posts and Telecommunications
[4]Xiangjiang Laboratory
[5]School of Computer Science and Engineering, Central South University
[6]Computer Science & Engineering, Konkuk University
[7]Department of Computer Science and Engineering, Seoul National University
xiongnanjin@szu.edu.cn, wzl446229@alibaba-inc.com, jpchen@bupt.edu.cn, yangliu@csu.edu.cn, bkoh@konkuk.ac.kr, seungwonh@snu.ac.kr, lijq@szu.edu.cn

Abstract

Entity alignment (EA) is crucial for integrating knowledge graphs (KGs) constructed from diverse sources. Conventional unsupervised EA approaches attempt to eliminate human intervention but often suffer from accuracy limitations. With the rise of large language models (LLMs), leveraging their capabilities for EA presents a promising direction. However, it introduces new challenges: formulating the LLM-based EA problem and extracting the background knowledge in LLMs to realize EA without human intervention. This paper proposes HLMEA, a novel hybrid language model-based unsupervised EA method. HLMEA formulates the EA task into a filtering and single-choice problem and synergistically integrates small language models (SLMs) and LLMs. Specifically, SLMs filter candidate entities based on textual representations generated from KG triples. Then, LLMs refine this selection to identify the most semantically aligned entities. An iterative self-training mechanism allows SLMs to distill knowledge from LLM outputs, enhancing the EA ability of hybrid language models in subsequent rounds cooperatively. We also conducted extensive experiments on benchmark datasets to evaluate HLMEA's performance. The results demonstrate that HLMEA significantly outperforms unsupervised and even supervised EA baselines, proving its potential for scalable and effective EA across large KGs. The code and data are available at https://github.com/xnjin-ai/HLMEA.

Introduction

Knowledge graphs (KGs) formulate human knowledge using a symbolic graph structure to enable machines to understand the semantics and logic. With the development of KG modeling and construction technologies, more and more KGs have been built in various domains to support KG-based applications such as question answering, reasoning, retrieval, and estimation (Wang et al. 2024b; Li et al. 2024; Agrawal et al. 2024; Chen et al. 2024; Liang et al. 2024; Liu, Wu, and Zhang 2024). However, the difficulty of inter-operation between KGs is ubiquitous due to the heterogeneous nature and incomplete data source. For instance, in various KGs, Apple's co-founder Steve Jobs might be identified by diverse relations and names, such as S. Jobs, Steve_Jobs, Jobs, or even appear in multiple languages.

Entity alignment (EA) aims to establish connections between KGs by identifying semantically equivalent entities. EA has attracted remarkable attention from researchers for enabling inter-communication and cooperation, i.e., KG fusion. Conventional EA work embeds symbolic KGs into vector spaces and then measures the similarities via entity embedding distances under the supervision of human labels (Yang et al. 2019; Tang et al. 2021; Chen et al. 2022). Semi-supervised EA methods are designed to reduce the dependency on high-cost human labels based on techniques like bootstrapping and iterative training (Sun et al. 2018; Xie et al. 2021). However, the above techniques still suffer from high-cost human annotation. Unsupervised EA has been studied to eliminate human intervention. Unimodal unsupervised approaches perform EA based on KG self-contained information by carefully designed mechanisms such as non-sampling calibration and relative similarity metric (Li and Song 2022; Liu et al. 2022). Some researchers incorporate auxiliary data and utilize multimodal pre-trained models to facilitate the EA accuracy (Liu et al. 2021; Chen et al. 2023).

With the emergence of Large Language Models (LLMs) such as GPT-4 (OpenAI 2023), traditional information technologies face revolutionary change. LLM-based EA is a promising direction to leverage the power of LLMs to improve EA performance further. However, LLM-based EA faces two challenges. The *first challenge* is how to formulate the EA problem to fit LLMs. A straightforward way is to compose a prompt containing a source entity and all target entities and feed it into an LLM to identify an aligned target entity. However, an entity typically has hundreds or thousands of triples, and a KG can have millions of entities. Simply injecting all entity information will result in an extremely long prompt that may exceed the context length limit or require an unacceptable long inference time.

LLMs are pre-trained on extensive and varied text data, which enables them to acquire a broad range of human knowledge and even surpass human performance in specific NLP tasks, such as summarization. However, when it

*Corresponding Author

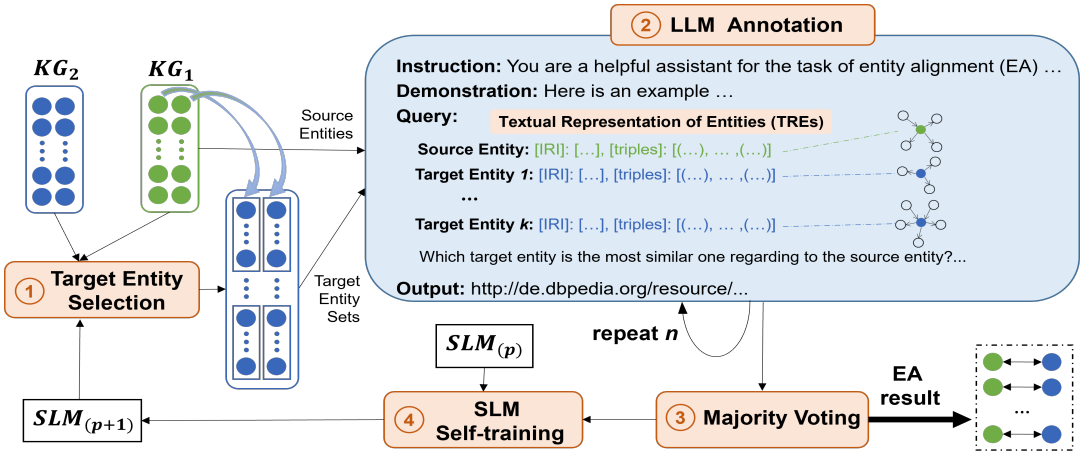


Figure 1: Architecture of the proposed unsupervised entity alignment (EA) method HLMEA. $\text{SLM}_{(p)}$ represents the fine-tuned SLM at round p. $\text{SLM}_{(0)}$ means the original version of SLM.

comes to the EA task, it is impractical to compare every entity pair and evaluate similarity scores using LLMs due to the high computational costs and hallucination issues (Ji et al. 2023). Moreover, in the unsupervised setting that is actively studied, there are no human labels for performing supervised fine-tuning on LLMs for EA. Therefore, the *second challenge* pertains to extracting background knowledge from LLMs to address the unsupervised EA task.

In this paper, we propose a novel **H**ybrid **L**anguage **M**odel-based unsupervised **EA** solution **HLMEA**. To address the first challenge, we formulate the EA problem into a filtering and single-choice problem. Specifically, a format of textual representations of entities (TREs) is proposed to describe entities concisely with less information loss. Then, given a source entity se and TREs as inputs, SLMs are adopted to filter the most promising top-k target entities tes based on the pre-trained embeddings to further reduce search space and prompt length. SLMs indicate the relatively small and lightweight pre-trained language models such as BERT (Devlin et al. 2019). Afterward, LLMs select an aligned entity from the filtered tes.

To address the second challenge, HLMEA enables LLMs and SLMs to cooperate iteratively and synergistically for the EA task. In detail, we propose an SLM self-training to fine-tune SLMs using the training data created based on LLM outputs. LLM selection (i.e., annotation) is repeated n times, and a majority voting is applied to reduce the uncertainty and instability of the LLM outputs. The training data generation aims to make the tes with higher SLM-measured similarity get more votes. The overall architecture of HLMEA is shown in Figrue 1. HLMEA iterates multiple rounds, and each round executes steps 1 to 4 in sequential order. In this way, SLMs distill the background knowledge from LLMs for better measuring entity similarities, and in turn, LLMs benefit from the improved quality of SLM-filtered tes.

Recently, an LLM-based EA approach ChatEA (Jiang et al. 2024) is presented. ChatEA proposes KG-code translation to understand KGs and enhance LLM's contextual knowledge. Embedding-based EA methods such as SimpleHHEA (Jiang et al. 2023) are employed as the backbone to filter candidates and perform reasoning and rethinking based on LLMs to identify aligned entities. However, ChatEA requires large-scale LLMs and extra entity descriptions to achieve high EA performance. Our distinction, without such dependence, is critical, as it allows our HLMEA to execute accurate EA using a smaller model like Qwen-7B (Bai et al. 2023) without auxiliary information. Our approach is, therefore, less resource-intensive and more flexible, particularly in scenarios lacking high-quality descriptive data. Overall, the main contributions are summarized as follows:

- We formalize the EA task into a filtering and single-choice problem to leverage the power of LLMs for improving accuracy. TRE is proposed to describe symbolic entities in an expressive and short textual format. Based on TREs, SLMs filter candidates by measuring similarities according to embedding distances to reduce the LLM search space and prompt size. Then, LLMs are employed to identify aligned entities from SLM-filtered candidates.
- We propose an unsupervised EA framework, HLMEA, that enables effective interaction and cooperation between LLMs and SLMs to enhance EA performance. HLMEA facilitates SLMs' distillation of background knowledge from LLMs through carefully designed self-training, and LLMs can take advantage of the improved quality of filtered candidates round by round.
- We conduct extensive experiments on benchmark datasets, and the results demonstrate the HLMEA's superiority in terms of effectiveness and scalability compared to unsupervised and supervised baselines.

Related Work

Supervised Entity Alignment. Several approaches have been proposed to address the problem of EA effectively, most of which are based on human labels. By leveraging ℓ_1 distance, HMAN (Yang et al. 2019) searches for the equivalent entities on the shared vector space after knowledge embedding. BERT-INT (Tang et al. 2021) defines a pairwise

margin loss to fine-tune a BERT model and then adopts an interactive method to embed KG for obtaining the aligned entity pairs. MSNEA (Chen et al. 2022) proposes the multimodal knowledge embedding and contrastive learning approach to achieve inter-modal enhancement fusion. Meanwhile, semi-supervised EA methods are developed to reduce the required human labels. Based on a bootstrapping method, BootEA (Sun et al. 2018) enriches the training data by iteratively labeling possible alignment. SEA (Pei et al. 2019) adopts a degree-aware KG embedding and defines a cycled consistent loss to incorporate unlabeled entities. Grounded in a dual-gated graph attention network, DuGa-DIT (Xie et al. 2021) represents an iterative training to update the attention scores by adding a new batch of labels. However, (semi-)supervised EA approaches suffer from human label collection, a high-cost and time-consuming process that limits the applicability and scalability.

Unsupervised Entity Alignment. Unsupervised EA has been actively investigated to eliminate human intervention. Based on the non-sampling calibration and gradual enhancement, UPLR (Li and Song 2022) reduces the influence of noisy labels and corrects the goal-oriented uncertainty. SelfKG (Liu et al. 2022) proposes a relative similarity metric to push negative alignments far away to bypass the supervision of positive pairs. SLOTAlign (Tang et al. 2023) transforms EA into an optimal transport problem and then jointly performs structure learning and optimal transport alignment. Some researchers infuse auxiliary information such as images and descriptions to further improve unsupervised EA's accuracy. EVA (Liu et al. 2021) leverages images as pivots and adopts an attention-based modality weighting scheme to fuse multimodal information for EA. Founded on graph neural networks, XGEA (Xu, Xu, and Su 2023) integrates multimodal information with structural details to enhance the alignment of entities. MEAformer (Chen et al. 2023) is a multimodal EA transformer approach that dynamically predicts mutual correlation coefficients among modalities for fine-grained entity-level modality fusion and alignment. In this paper, we propose to realize unsupervised EA without any auxiliary data by formulating EA into a filtering and single-choice problem. Then, the ability of LLMs and SLMs is mutually integrated and reinforced for the EA task.

Problem Definition

We denote a knowledge graph by $\mathcal{KG} = (\mathcal{S}, \mathcal{P}, \mathcal{O})$, where $\mathcal{S}$, $\mathcal{P}$, and $\mathcal{O}$ indicate collections of subjects, predicates, and objects. Each subject denotes an entity, and an object indicates an entity or attribute. A predicate connects two entities as a relation or links an entity and an attribute as a property. EA aims to find identical entities in two different KGs. The EA problem is formally defined as follows:

Definition 1 (Entity Alignment: EA) Given a source and a target knowledge graph $\mathcal{KG}_s$ and $\mathcal{KG}_t$, EA aims to work out a list of aligned entity pairs such that $\mathcal{A}_{\mathcal{KG}_s,\mathcal{KG}_t} = \{(se, te) \in \mathcal{E}_s \times \mathcal{E}_t, \mathcal{E}_s \in \mathcal{KG}_s, \mathcal{E}_t \in \mathcal{KG}_t | se \sim te\}$, where $\sim$ and $\mathcal{E}$ indicate an equivalence relation and a set of entities.

In this paper, we aim to address the EA problem by exploiting the background knowledge contained in large language models (LLMs) without human intervention. The LLM-based EA problem is formalized as follows:

Definition 2 (LLM-based Entity Alignment: LEA) Given a source entity $se \in \mathcal{KG}_s$ and target entities $te \in \mathcal{KG}_t$, the goal of LEA is to select a most similar te by leveraging the power of LLMs through a carefully designed prompt. The prompt typically contains instructions describing the task and symbolic-semantic information about se and te.

Proposed Approach

This section introduces the proposed hybrid language model-based unsupervised entity alignment approach HLMEA. As shown in Figure 1, HLMEA consists of four modules: target entity selection, LLM annotation, majority voting, and SLM self-training. In round 0 step 1, target entity selection chooses top-k candidate target entities tes regarding a given source entity se based on an SLM. Then, in step 2, LLM annotation creates a prompt composed of the instruction, demonstration, and query and sends it to an LLM to identify a most similar te. In step 3, majority voting repeats the LLM annotation n times and aggregates the results to reduce the uncertainty caused by the LLM hallucinations (Zhang et al. 2023). In step 4, SLM self-training generates training data based on the LLM output distributions and fine-tunes the SLM. Afterward, HLMEA iterates steps 1 to 4 at each round and executes r rounds. In this way, hybrid language models interact and cooperate to improve performance synergistically.

Target Entity Selection

Given an se, tes are typically the entire target KG entities $\mathcal{E}_t \in \mathcal{KG}_t$ in the EA task. However, the LLM input context limitation makes it difficult to put all the $\mathcal{E}_t$ information into a prompt, especially when $\mathcal{KG}_t$ is large. Although infinite context transformers have been proposed recently (Munkhdalai, Faruqui, and Gopal 2024), the issues of enormous inference cost and GPU memory usage still exist. Hence, target entity selection is designed to choose top-k similar tes to reduce the number of candidates and prompt length based on SLMs. In this way, LLMs infuse the SLM's similarity-measuring ability to perform LEA annotation.

Firstly, symbolic representations of entities are transformed into text to feed into an SLM to measure the similarities between entities. Intuitively, the textual representation of entities (TRE) can be generated by simply aggregating the related attributes and relation triples. However, an entity can have hundreds or thousands of triples, which raises the need for text pruning. Therefore, we selected m triples from each attribute, relation-out, and relation-in triples. The relation triples are categorized into relation-out and relation-in subsets to ensure a balanced selection. This division prevents the bias of exclusively picking relation triples, solely outgoing or incoming. The predicate guides the triple selection process to provide valuable identifying information among various entities. Inspired by TF-IDF (Sparck Jones 1972), predicate scores are computed as follows:

$$PF(p, e) = |triple_{(e)}(p)|/|triple_{(e)}(\cdot)| \quad (1)$$

IRI	http://dbpedia.org/resource/Rick_Rubin
Attribute Triples	(Rick_Rubin, name, Rick Rubin), ..., (Rick_Rubin, birthName, Frederick Jay Rubin@en)
Relation -out Triples	(Rick_Rubin, birthPlace, United_States), ..., (Rick_Rubin, occupation, Record_producer)
Relation -in Triples	(Run–D.M.C., associatedActs, Rick_Rubin), ..., (Def_Jam_Recordings, founder, Rick_Rubin)

Table 1: Textual Representation of an Entity (TRE).

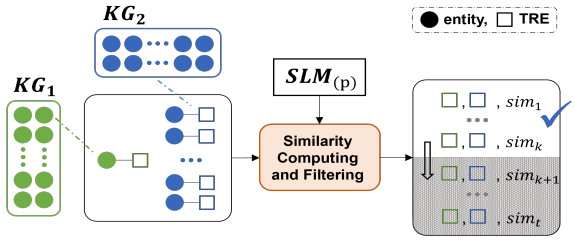


Figure 2: Target Entity Selection based on TREs and SLM.

$$IKF(p, \mathcal{KG}) = |entity_{(\mathcal{KG})}(p)|/|entity_{(\mathcal{KG})}(\cdot)| \quad (2)$$

$$PF-IKF(p, \mathcal{KG}) = PF(p, e) \times IKF(p, \mathcal{KG}) \quad (3)$$

where $triple_{(e)}(p)$ denote the triples of an entity e that have predicate p, and $entity_{(\mathcal{KG})}(p)$ indicates the entities of a KG that have triples containing p. PF($\cdot$) and IKF($\cdot$) measure the importance of p within a certain e and the universality among a KG. Top-m predicates are selected based on the PF-IKF scores. One triple is randomly picked for each predicate if multiple triples exist. Finally, the KG domain (e.g., http://dbpedia.org/resource/) is abbreviated to reduce text size further. This reduction would not cause information loss since the domain and name are described in the entity IRI, and all the entities in a single KG commonly share the same domain prefix. Table 1 illustrates an example of TRE.

Secondly, top-k candidate tes are retrieved based on TREs and an SLM, as shown in Figure 2. Entity embeddings are inferred using the TREs based on the SLM. The pairwise similarities $sim(se, te_1), ..., sim(se, te_t)$ are measured via the central moment discrepancy (CMD) (Zellinger et al. 2022), which is a distance metric that calculates the similarity by quantifying the discrepancy in central moments. We set variables a and b as 0 and 1 for simplicity and a smaller CMD value indicates a higher similarity between two entity embeddings. Then, tes are sorted by the descending order of similarities, and top-k tes are chosen as the candidates to be further distinguished by LLMs.

LLM Annotation and Majority Voting

LLM annotation identifies entity alignments from filtered tes based on LLMs through carefully designed prompts. A prompt consists of an instruction and a query. The instruction describes the EA task as a single-choice problem and expresses the knowledge triples. The query contains TREs of an se and k tes. Since the LLM outputs are unstable due to the hallucinations issue, LLM annotation is performed n times to reduce the uncertainty, and majority voting is employed to select representative results under the principle of plurality. Table 2 shows examples of LLM output distribution. The numbers indicate hit counts, and the underlined tes denote the LEA results after majority voting. If multiple top tes are associated with the same hit counts, then te with the smallest rank is picked to reflect the choice of SLM.

http://dbpedia.org/resource/Frank_Simek:
① (http://de.dbpedia.org/.../Frank_Simek, 4),
② (http://de.dbpedia.org/.../Will_Lee_Musiker, 0),
③ (http://de.dbpedia.org/.../Frank_Langella, 0),
④ (http://de.dbpedia.org/.../Frank_Wiblishauser, 1),
⑤ (http://de.dbpedia.org/.../Frank_Gatski, 0)
...
http://dbpedia.org/resource/Star_Trek:_Deep_Space_Nine:
① (http://de.dbpedia.org/.../Space_2063, 0),
② (http://de.dbpedia.org/.../Starship_Troopers_(Film), 0),
③ (http://de.dbpedia.org/..._Deep_Space_Nine, 5),
④ (http://de.dbpedia.org/.../Star_Wars:_Battlefront_II, 0),
⑤ (http://de.dbpedia.org/.../Marvin_Pourie, 0)

Table 2: An example of LLM output distribution. The circled numbers denote the ranks of tes regarding SLM similarity. The lower rank means the higher similarity.

SLM Self-Training

An SLM is fine-tuned at each round to distill knowledge from LLMs and improve the effectiveness of SLM entity embedding. The self-training data with the form of $(TRE_{se_i}, TRE_{pos_i}, TRE_{neg_i})$ is generated based on the LLM output distributions. $TRE_{se_i}, TRE_{pos_i}, TRE_{neg_i}$ denote TRE of i^{th} se, positive sample, and negative sample. Positive samples are the results of majority voting. The rationale behind the negative sampling is that tes assigned with greater similarity by the SLM should receive more hits from the LLM annotation. Specifically, tes are sorted in descending order regarding hit counts. Then, we compare the hit count order and SLM similarity order, and the first order-violating te with the smallest similarity rank, except the positive sample, is chosen as the negative sample.

For instance, given an se shown in Table 2 http://dbpedia.org/resource/Frank_Simek, the tes, ranks of SLM similarity, and ranks of hit counts are: Frank_Simek-①-①, Will_Lee_(Musiker)-②-③, Frank_Langella-③-④, Frank_Wiblishauser-④-②, and Frank_Gatski-⑤-⑤. Therefore, the first order-violating Will_Lee_(Musiker) is selected as the negative sample to push its embeddings far away from se embeddings. Finally, the constructed training data is utilized to fine-tune $SLM_{(p)}$ to generate $SLM_{(p+1)}$ for next-round usage by using a pairwise margin-based loss.

Experiments

We have conducted experiments on benchmark datasets to verify the effectiveness of the proposed HLMEA. The research questions that we aim to answer are listed as follows:

Category	Unsupervised EA method	Configurations			DBP15K$_{ZH-EN}$		DBP15K$_{JA-EN}$		DBP15K$_{FR-EN}$		Average
		Stru.	Attr.	Aux.	Hit@1	MRR	Hit@1	MRR	Hit@1	MRR	Hit@1
Uni-modal	MultiKE (Zhang et al. 2019)	✓	✓	×	.509	-	.393	-	.639	-	.514
	SelfKG (Liu et al. 2022)	✓	×	×	.745	-	.816	-	.957	-	.839
	UPLR (Li and Song 2022)	✓	×	×	.902	.927	.912	.937	.967	.974	.927
	SLOTAlign (Tang et al. 2023)	✓	✓	×	.890	-	.930	-	**.992**	-	.937
	HLMEA (ours)	✓	✓	×	**.930**	**.934**	**.938**	**.950**	.986	**.989**	**.951**
	– improvement				**+2.8%**	**+0.7%**	**+0.8%**	**+1.3%**	**-0.6%**	**+1.5%**	**+1.4%**
Multi-modal	EVA (Liu et al. 2021)	✓	✓	$\mathcal{I}$	.752	.804	.737	.791	.731	.792	.740
	MCLEA (Lin et al. 2022)	✓	✓	$\mathcal{I}$	.803	.851	.781	.838	.780	.838	.788
	ICLEA (Zeng et al. 2022)	✓	✓	$\mathcal{D}$	.884	-	.919	-	**.986**	-	.930
	XGEA (Xu, Xu, and Su 2023)	✓	✓	$\mathcal{I}$	.823	.865	.811	.859	.826	.874	.820
	MEAformer (Chen et al. 2023)	✓	✓	$\mathcal{I}$	**.909**	**.933**	**.950**	**.965**	.972	**.983**	**.944**

Table 3: Unsupervised EA results on the bi-lingual dataset DBP15K regarding Hit@1 and MRR. EA methods are categorized into unimodal and multimodal according to whether auxiliary information is incorporated. Bold and underlined numbers indicate their corresponding category's best and runner-up scores. $\mathcal{I}$ and $\mathcal{D}$ represent images and descriptions.

- (Q1) Does HLMEA outperform state-of-the-art (SOTA) unsupervised EA methods on the benchmark datasets?
- (Q2) Does HLMEA effectively compress entity information to fit LLMs without harming EA performance?
- (Q3) Does the cooperation between the LLM and SLM contribute to the EA performance synergistically?
- (Q4) How essential are different HLMEA components and variables regarding the EA task?

Experimental Settings

Datasets. We employ nine widely used EA datasets DBP15K$_{ZH-EN,JA-EN,FR-EN}$ (Sun, Hu, and Li 2017), DBP15K$_{DE-EN,FR-EN}$, DW15K V1, DY15K V1, and DBP100K$_{DE-EN,FR-EN}$ (Sun et al. 2020). DW15K V1 and DY15K V1 are monolingual and cross-KG datasets written in English, and others are bi-lingual and single-KG datasets constructed to test the EA ability for entities of different languages. DBP100K$_{DE-EN,FR-EN}$ V1 are large-scale datasets used to verify the scalability of EA methods, which contain 100,000 aligned entity pairs.

Baselines. We compared our HLMEA with supervised and unsupervised EA baselines. Supervised baselines train the EA models based on human labels. Unsupervised baselines execute EA without human annotation, and HLMEA belongs to the unsupervised category.

Evaluation Metrics. Following the previous works (Liu et al. 2023; Chen et al. 2023), Hit@k (k = 1, 3, 5, 10, 20) and mean reciprocal rank (MRR) are adopted as metrics to evaluate the effectiveness of HLMEA. Besides, LLM accuracy (LA) is employed to measure the identification ability among SLM selected top-k target entities, which is formalized as LA = Hit@1/Hit@k.

Experimental Environment. We implemented HLMEA using Python (version 3.9.18) with a PyTorch (version 2.1.2) backend. Experiments were conducted on a server running Ubuntu 22.04, which has an AMD Ryzen 9 7950X Processor, 128GB memory, and an NVIDIA RTX A6000 GPU.

Parameter Settings. LLMs adopted were large-scale close-sourced ChatGPT (OpenAI 2022) (version gpt-3.5-turbo-1106) and ERNIE (BaiduResearch 2023) (version ERNIE-3.5-8K-0329), and relatively small and open-sourced Qwen (Bai et al. 2023) (version 7B). Pre-trained multi-lingual language models E5 (Wang et al. 2024a), LaBSE (Feng et al. 2022), MPNet and MiniLM (Reimers and Gurevych 2019) were employed as SLMs. AdamW (Loshchilov and Hutter 2019) was adopted as the optimizer, and the learning rate was set as 1e^{-5}. To reduce the utilization of LLMs, 20% of data was employed to train the SLMs with the self-supervision of LLM annotation, then the HLMEA inference results on the remaining 80% were reported. LLM annotation repetition n and candidate entity quantity top-k were set in the range [3, 20].

(Q1) Main Results on Benchmark Datasets

The evaluation results on DBP15K with a comparison of unimodal and multimodal unsupervised EA methods are shown in Table 3. The multimodal methods incorporate auxiliary information such as images and descriptions. Unimodal methods utilize merely dataset self-contained information, like our HLMEA. UPLR and SLOTAlign realized decent performance in the unimodal category based on techniques such as non-sampling calibration and multi-view structure learning. Multimodal methods outperformed existing unimodal approaches due to the complementary utilization of external information. Our HLMEA, equipped with ERNIE and LaBSE, achieved the best performance among unimodal methods in almost all cases and represented a superior average Hit@1 score even compared to multimodal EA methods. It was because HLMEA formulates the filtering and single-choice problem for synergistically comprising LLM and SLM abilities. Multimodal MEAformer recorded the best scores on DBP15K$_{JA-EN}$ dataset. It may be because high-quality images complemented the difficulty caused by the language grammar difference.

Table 4 shows the EA results on DBP15K, DW15K, and DY15K V1 datasets. The supervised EA methods employing 20% seeds (human labels) for training were selected as baselines since no unsupervised records were reported to the best of our knowledge. RDGCN represented runner-up scores by modeling the attentive interactions between the primary and dual relation graphs. The proposed HLMEA

EA Method	Seed	DBP15K$_{DE-EN}$ V1		DBP15K$_{FR-EN}$ V1		DW15K V1		DY15K V1		Average
		Hit@1	MRR	Hit@1	MRR	Hit@1	MRR	Hit@1	MRR	Hit@1
BootEA (Sun et al. 2018)	20%	.675	.740	.507	.603	.572	.649	.739	.788	.623
RDGCN (Wu et al. 2019)	20%	.830	.859	.755	.800	.515	.584	.931	.949	.758
GAEA (Xie et al. 2023)	20%	.684	.760	.486	.602	.562	**.654**	.608	.688	.585
RHGN (Liu et al. 2023)	20%	.704	.771	.500	.603	.560	.644	.708	.762	.618
HLMEA(ours)	**0**	**.955**	**.965**	**.957**	**.968**	**.618**	.634	**.967**	**.981**	**.874**
– improvement		**+12.5%**	**+10.6%**	**+20.2%**	**+16.8%**	**+4.6%**	**-2.0%**	**+3.6%**	**+3.2%**	**+11.6%**

Table 4: Results on DBP15K, DW15K, and DY15K V1 datasets. Bold and underlined numbers indicate the best and runner-ups.

EA Method	Seed	DBP100K$_{FR-EN}$ V1			DBP100K$_{DE-EN}$ V1			Average
		Hit@1	Hit@5	MRR	Hit@1	Hit@5	MRR	Hit@1
BootEA (Sun et al. 2018)	20%	.482	.515	.499	.518	.673	.592	0.500
RSN4EA (Guo, Sun, and Hu 2019)	20%	.293	.452	.371	.430	.570	.497	0.362
MultiKE (Zhang et al. 2019)	20%	.629	.680	.655	.668	.712	.690	0.649
RDGCN (Wu et al. 2019)	20%	.640	.732	.683	.722	.794	.756	0.681
HLMEA(ours)	**0**	**.925**	**.965**	**.940**	**.899**	**.932**	**.917**	**0.912**
– improvement		**+28.5%**	**+23.3%**	**+25.7%**	**+17.7%**	**+13.8%**	**+16.1%**	**+23.1%**

Table 5: EA results on DBP100K V1 datasets. Bold and underlined numbers indicate the best and runner-ups.

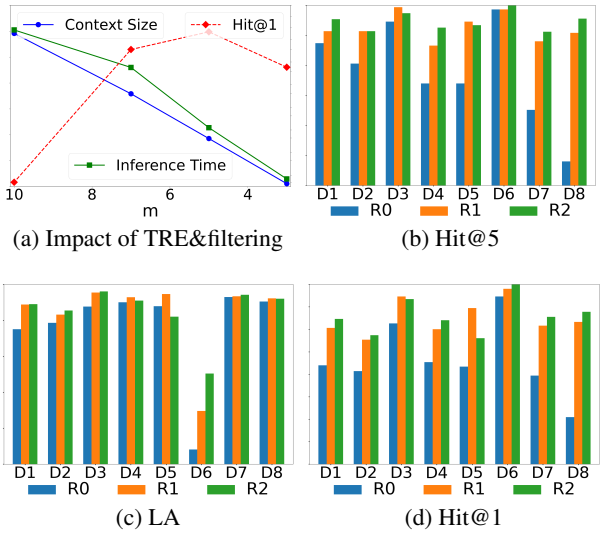
outperformed RDGCN by 11.6% in average Hit@1 without any seed. It is because HLMEA incorporates and adapts the background knowledge contained in LLMs and pre-trained linguistic knowledge of SLMs to the EA task. The gap grew to 23.1% on the DBP100K V1 datasets, shown in Table 5, demonstrating the scalability of HLMEA. LLM and SLM employed in Table 4 and 5 were ChatGPT and LaBSE.

(Q2) Impact of Entity Information Compression

HLMEA compresses entity information using TREs and SLM filtering, which reduces the number of triples to m for each entity and candidate tes to k. Figure 3(a) shows the HLMEA, equipped with Qwen and LaBSE, R0 results on the DY15K V1 dataset with varying m (k was set as the same as m). With smaller m, the LLM input context size and inference time for each se decreased spontaneously. On the contrary, EA performance increased and achieved the highest Hit@1 with $m = 5$. It demonstrates that HLMEA prunes entity information and forms proper prompts to efficiently exploit LLMs' ability for the EA task.

(Q3) Cooperation between LLMs and SLMs

To test the effectiveness of cooperation between LLMs and SLMs, the HLMEA performances of each round were recorded and shown in Figure 3. Hit@5 scores represent the SLM's ability to select high-quality candidates that include the ground truth. As shown in Figure 3(b), the SLM performance increased significantly from round 0 (R0) to round 1 (R1) and continued growing in most cases. It indicates that SLMs effectively distilled background knowledge from LLMs through the proposed SLM self-training. Figure 3(c) demonstrates the LLM's performance of identifying the ground truth from the top-k candidates under different rounds. LA scores rose from R0 to R1 and grew on most datasets since the quality of SLM-generated candidates increased. Moreover, this mechanism of cooperation resulted in the increasing EA performance with more rounds of execution, which is shown in Figure 3(d). The imperfection of SLM and LLM may produce noisy candidates or training data that may drop the EA performance in the later rounds.



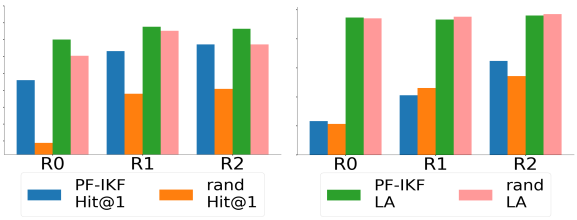
(a) Impact of TRE&filtering (b) Hit@5

(c) LA (d) Hit@1

Figure 3: Impact of TRE and SLM filtering (a), where ranges of context size, Hit@1, and inference time are [600, 2000] tokens, [0.8, 1], and [1, 2.6] seconds, and cooperation between LLM and SLM (b to d), where D1 to D8 denote DBP15K$_{ZH-EN}$, DBP15K$_{JA-EN}$, DBP15K$_{FR-EN}$, DBP15K$_{DE-EN}$ V1, DBP15K$_{FR-EN}$ V1, DY15K V1, DBP100K$_{DE-EN}$ V1, and DBP100K$_{FR-EN}$ V1.

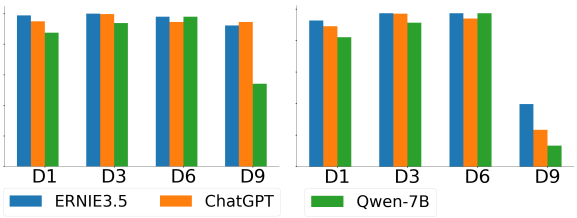
(Q4) Ablation Study

Ablation studies investigated how parameters, such as TRE generation strategy, LLM type, SLM type, and top-k, impact the EA performance. TRE generation strategies tested



(a) DBP15K$_{ZH-EN}$ (b) DW15K V1

Figure 4: Impact of $PF\text{-}IKF$ vs. random strategies, where the range of y-axis is [0, 1].



(a) LA (b) Hit@1

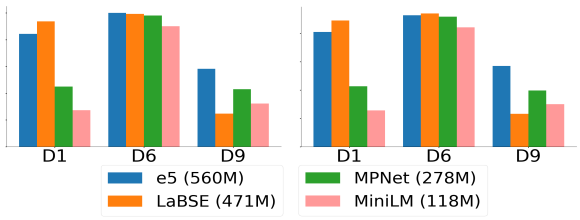
Figure 5: Performance with Different LLMs, where y-axis range is [0,1]. D1, D3, D6, and D9 denote DBP15K$_{ZH-EN}$, DBP15K$_{FR-EN}$, DY15K V1, and DW15K V1.

were $PF\text{-}IKF$ and $Rand$. In $PF\text{-}IKF$, m triples with the highest PF-IKF scores are chosen, and the $Rand$ strategy randomly selects m triples. Figure 4 shows the results with different TRE generation strategies on two datasets. LLM and SLM were set as ChatGPT and LaBSE. $PF\text{-}IKF$ outperformed $Rand$ on the DBP15K$_{ZH-EN}$ dataset by a large margin in terms of Hit@1 and LA. It denotes that $PF\text{-}IKF$ scores provide valuable guidance to select triples for measuring the similarities of entities. On the DW15K V1 dataset, $PF\text{-}IKF$ recorded higher Hit@1 at R2, but the gap was narrowed. This is because DBpedia and Wikidata have different distributions of triples and little overlap.

Figure 5 shows the HLMEA performance with different LLMs including closed ERNIE3.5 and ChatGPT, and open-source Qwen-7B. SLM was set as LaBSE. In general, closed LLMs represented better performance than the relatively small Qwen. However, Qwen scored higher than ChatGPT on the DW15K V1 dataset. Note that open-source LLMs can be deployed locally and optimized by fine-tuning.

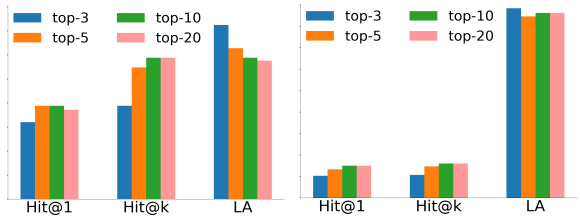
The performance of HLMEA with ChatGPT using different multi-lingual SLMs is illustrated in Figure 6. e5 (MiniLM) is the biggest (smallest) SLM with 560M (118M) parameters. In general, the performance is proportionate to the scale of parameters. As an exception, the second-largest LaBSE scored the best on two datasets and the worst on DW15K V1. It denotes that over 300M SLMs are competent in EA, and one SLM may not perform the best universally.

Figure 7 demonstrates the performance of HLMEA with ChatGPT, LaBSE, and varying top-k, which means the number of candidates to be selected by an SLM. Hit@k corresponds to top-k, for example Hit@10 was set for top-10. With growing k, the ability of SLMs (LLMs) increased with expanding candidates. Overall Hit@1 rose since the expanding ratio of Hit@5 was higher than that of LA decreasing ratio. However, there was no improvement from 10 to 20, and the greater k would result in a larger prompt context length.



(a) Hit@5 (b) Hit@1

Figure 6: Performance with Different SLMs, where range of y-axis is [0,1].



(a) DBP15K$_{ZH-EN}$ (b) DW15K V1

Figure 7: Performance with Varying top-k.

Discussion on the Cost

The proposed HLMEA relies solely on the knowledge embedded in LLMs and SLMs without human intervention, making it a cost-effective option for annotation in real-world EA scenarios. According to the previous work (Wang et al. 2021), human annotation costs \$0.11 per 50 tokens and an EA example containing approximately 1,000 tokens, which results in \$2.2 per EA labeling. HLMEA typically converges after three rounds of execution, each round with three repetitions. Using ChatGPT as an example, which charges \$1 per 1 million tokens for version 3.5-turbo-1106, the cost of using HLMEA for one EA example is calculated as \$1×1,000×3×3/1,000,000=\$0.009. It demonstrates that HLMEA offers a significantly cheaper alternative for EA annotation while maintaining satisfactory performance.

Conclusion

This paper, we propose a novel unsupervised EA framework, HLMEA, that facilitates the cooperation between LLMs and SLMs to improve EA performance synergistically. Comparative experiments and ablation studies were conducted on benchmark datasets, and the results validate that HLMEA outperforms baselines in terms of effectiveness and scalability. We hope this work will inspire further research into incorporating LLMs for more downstream tasks.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62306287, 62073225, 62203134, 62172451), National Natural Science Funds for Distinguished Young Scholar (No. 62325307), Zhejiang Provincial Natural Science Foundation of China (No. LY23F020012), Natural Science Foundation of Guangdong Province (No. 2023B1515120038), Guangdong "Pearl River Talent Recruitment Program" (No. 2019ZT08X603), Guangdong "Pearl River Talent Plan" (No. 2019JC01X235), Shenzhen Science and Technology Innovation Commission (No. KJZD20230923113801004, 20220809141216003), Beijing Natural Science Foundation(No. L233034), Scientific Instrument Developing Project of Shenzhen University (No. 2023YQ019), Open Project of Xiangjiang Laboratory (No. 23XJ03006), and National Research Foundation of Korea Grant Funded by the Korea Government (MIST) (No. RS-2024-00457435).

References

Agrawal, G.; Pal, K.; Deng, Y.; Liu, H.; and Chen, Y.-C. 2024. CyberQ: Generating Questions and Answers for Cybersecurity Education Using Knowledge Graph-Augmented LLMs. In *AAAI Conf. on Artificial Intelligence (AAAI)*, 23164–23172.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

BaiduResearch. 2023. Introducing ERNIE 3.5. Accessed: 2024-07-23. Http://research.baidu.com/Blog/index-view?id=185.

Chen, L.; Li, Z.; Xu, T.; Wu, H.; Wang, Z.; Yuan, N. J.; and Chen, E. 2022. Multi-modal siamese network for entity alignment. In *ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining (KDD)*, 118–126.

Chen, Z.; Chen, J.; Zhang, W.; Guo, L.; Fang, Y.; Huang, Y.; Zhang, Y.; Geng, Y.; Pan, J. Z.; Song, W.; et al. 2023. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid. In *ACM Int'l Conf. on Multimedia (MM)*, 3317–3327.

Chen, Z.; Zhou, Q.; Shen, Y.; Hong, Y.; Sun, Z.; Gutfreund, D.; and Gan, C. 2024. Visual Chain-of-Thought Prompting for Knowledge-Based Visual Reasoning. In *AAAI Conf. on Artificial Intelligence (AAAI)*, 1254–1262.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 4171–4186.

Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; and Wang, W. 2022. Language-agnostic BERT Sentence Embedding. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 878–891.

Guo, L.; Sun, Z.; and Hu, W. 2019. Learning to exploit long-term relational dependencies in knowledge graphs. In *Int'l Conf. on Machine Learning (ICML)*, 2505–2514. PMLR.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Jiang, X.; Shen, Y.; Shi, Z.; Xu, C.; Li, W.; Li, Z.; Guo, J.; Shen, H.; and Wang, Y. 2024. Unlocking the Power of Large Language Models for Entity Alignment. *arXiv preprint arXiv:2402.15048*.

Jiang, X.; Xu, C.; Shen, Y.; Su, F.; Wang, Y.; Sun, F.; Li, Z.; and Shen, H. 2023. Rethinking gnn-based entity alignment on heterogeneous knowledge graphs: New datasets and a new method. *CoRR*.

Li, J.; and Song, D. 2022. Uncertainty-aware Pseudo Label Refinery for Entity Alignment. In *The Web Conf. (WWW)*, 829–837.

Li, Z.; Fan, S.; Gu, Y.; Li, X.; Duan, Z.; Dong, B.; Liu, N.; and Wang, J. 2024. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. In *AAAI Conf. on Artificial Intelligence (AAAI)*, 18608–18616.

Liang, M.; Du, J.; Liang, Z.; Xing, Y.; Huang, W.; and Xue, Z. 2024. Self-Supervised Multi-Modal Knowledge Graph Contrastive Hashing for Cross-Modal Search. In *AAAI Conf. on Artificial Intelligence (AAAI)*, 13744–13753.

Lin, Z.; Zhang, Z.; Wang, M.; Shi, Y.; Wu, X.; and Zheng, Y. 2022. Multi-modal contrastive representation learning for entity alignment. In *Int'l Conf. on Computational Linguistics (COLING)*, 2572–2584.

Liu, F.; Chen, M.; Roth, D.; and Collier, N. 2021. Visual pivoting for (unsupervised) entity alignment. In *AAAI Conf. on Artificial Intelligence (AAAI)*, 4257–4266.

Liu, R.; Wu, L.; and Zhang, P. 2024. KG-TREAT: Pre-training for Treatment Effect Estimation by Synergizing Patient Data with Knowledge Graphs. In *AAAI Conf. on Artificial Intelligence (AAAI)*, 8805–8814.

Liu, X.; Hong, H.; Wang, X.; Chen, Z.; Kharlamov, E.; Dong, Y.; and Tang, J. 2022. Selfkg: Self-supervised entity alignment in knowledge graphs. In *ACM Web Conference (WWW)*, 860–870.

Liu, X.; Zhang, K.; Liu, Y.; Chen, E.; Huang, Z.; Yue, L.; and Yan, J. 2023. RHGN: Relation-gated Heterogeneous Graph Network for Entity Alignment in Knowledge Graphs. In *Findings of the Association for Computational Linguistics (ACL)*, 8683–8696.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *Int'l Conf. on Learning Representations (ICLR)*.

Munkhdalai, T.; Faruqui, M.; and Gopal, S. 2024. Leave No Context Behind: Efficient Infinite Context Transformers with Infini-attention. arXiv:2404.07143.

OpenAI. 2022. Introducing ChatGPT. Accessed: 2024-07-23. Https://openai.com/index/chatgpt/.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Pei, S.; Yu, L.; Hoehndorf, R.; and Zhang, X. 2019. Semi-supervised entity alignment via knowledge graph embedding with awareness of degree difference. In *The Web Conf. (WWW)*, 3130–3136.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conf. on Empirical Methods in Natural Language Processing and Int'l Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992.

Sparck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1): 11–21.

Sun, Z.; Hu, W.; and Li, C. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. In *Int'l Semantic Web Conf. (ISWC)*, 628–644. Springer.

Sun, Z.; Hu, W.; Zhang, Q.; and Qu, Y. 2018. Bootstrapping entity alignment with knowledge graph embedding. In *Int'l Joint Conf. on Artificial Intelligence (IJCAI)*.

Sun, Z.; Zhang, Q.; Hu, W.; Wang, C.; Chen, M.; Akrami, F.; and Li, C. 2020. A benchmarking study of embedding-based entity alignment for knowledge graphs. *Proceedings of the VLDB Endowment*, 13(12): 2326–2340.

Tang, J.; Zhang, W.; Li, J.; Zhao, K.; Tsung, F.; and Li, J. 2023. Robust attributed graph alignment via joint structure learning and optimal transport. In *IEEE Int'l Conf. on Data Engineering (ICDE)*, 1638–1651. IEEE.

Tang, X.; Zhang, J.; Chen, B.; Yang, Y.; Chen, H.; and Li, C. 2021. BERT-INT: a BERT-based interaction model for knowledge graph alignment. In *Int'l Joint Conf. on Artificial Intelligence (IJCAI)*, 3174–3180.

Wang, L.; Yang, N.; Huang, X.; Yang, L.; Majumder, R.; and Wei, F. 2024a. Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint arXiv:2402.05672*.

Wang, S.; Liu, Y.; Xu, Y.; Zhu, C.; and Zeng, M. 2021. Want To Reduce Labeling Cost? GPT-3 Can Help. In *Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 4195–4205.

Wang, Y.; Lipka, N.; Rossi, R. A.; Siu, A.; Zhang, R.; and Derr, T. 2024b. Knowledge graph prompting for multi-document question answering. In *AAAI Conf. on Artificial Intelligence (AAAI)*, 19206–19214.

Wu, Y.; Liu, X.; Feng, Y.; Wang, Z.; Yan, R.; and Zhao, D. 2019. Relation-aware entity alignment for heterogeneous knowledge graphs. In *Int'l Joint Conf. on Artificial Intelligence (IJCAI)*.

Xie, F.; Zeng, X.; Zhou, B.; and Tan, Y. 2023. Improving knowledge graph entity alignment with graph augmentation. In *Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD)*, 3–14. Springer.

Xie, Z.; Zhu, R.; Zhao, K.; Liu, J.; Zhou, G.; and Huang, J. X. 2021. Dual gated graph attention networks with dynamic iterative training for cross-lingual entity alignment. *ACM Transactions on Information Systems*, 40(3): 1–30.

Xu, B.; Xu, C.; and Su, B. 2023. Cross-modal graph attention network for entity alignment. In *ACM Int'l Conf. on Multimedia (MM)*, 3715–3723.

Yang, H.-W.; Zou, Y.; Shi, P.; Lu, W.; Lin, J.; and Sun, X. 2019. Aligning cross-lingual entities with multi-aspect information. In *Conf. on Empirical Methods in Natural Language Processing and Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, 4430–4440.

Zellinger, W.; Grubinger, T.; Lughofer, E.; Natschläger, T.; and Saminger-Platz, S. 2022. Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning. In *Int'l Conf. on Learning Representations (ICLR)*.

Zeng, K.; Dong, Z.; Hou, L.; Cao, Y.; Hu, M.; Yu, J.; Lv, X.; Cao, L.; Wang, X.; Liu, H.; et al. 2022. Interactive contrastive learning for self-supervised entity alignment. In *ACM Int'l Conf. on Information & Knowledge Management (CIKM)*, 2465–2475.

Zhang, M.; Press, O.; Merrill, W.; Liu, A.; and Smith, N. A. 2023. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*.

Zhang, Q.; Sun, Z.; Hu, W.; Chen, M.; Guo, L.; and Qu, Y. 2019. Multi-view knowledge graph embedding for entity alignment. In *Int'l Joint Conf. on Artificial Intelligence (IJCAI)*, 5429–5435.