

Knowledge Graph Completion with Relation-Aware Anchor Enhancement

Duanyang Yuan^{1*}, Sihang Zhou^{1*}, Xiaoshu Chen², Dong Wang¹, Ke Liang²,
Xinwang Liu², Jian Huang^{1†}

¹College of Intelligence Science and Technology, National University of Defense Technology, Changsha, China

²College of Computer Science and Technology, National University of Defense Technology, Changsha, China
{ydy_n1, shzhou, chenxs, lizexin, liangke22, xinwangliu, huang_jian}@nudt.edu.cn

Abstract

Text-based knowledge graph completion methods take advantage of pre-trained language models (PLM) to enhance intrinsic semantic connections of raw triplets with detailed text descriptions. Typical methods in this branch map an input query (textual descriptions associated with an entity and a relation) and its candidate entities into feature vectors, respectively, and then maximize the probability of valid triples. These methods are gaining promising performance and increasing attention for the rapid development of large language models. According to the property of the language models, the more related and specific context information the input query provides, the more discriminative the resultant embedding will be. In this paper, through observation and validation, we find a neglected fact that the relation-aware neighbors of the head entities in queries could act as effective contexts for more precise link prediction. Driven by this finding, we propose a relation-aware anchor enhanced knowledge graph completion method (RAA-KGC). Specifically, in our method, to provide a reference of what might the target entity be like, we first generate anchor entities within the relation-aware neighborhood of the head entity. Then, by pulling the query embedding towards the neighborhoods of the anchors, it is tuned to be more discriminative for target entity matching. The results of our extensive experiments not only validate the efficacy of RAA-KGC but also reveal that by integrating our relation-aware anchor enhancement strategy, the performance of current leading methods can be notably enhanced without substantial modifications.

Code — <https://github.com/DayanaYuan/RAA-KGC>

Introduction

A knowledge graph is a structured semantic knowledge base that stores world knowledge through triples, i.e., (*head entity*, *relation*, *tail entity*), or (*h*, *r*, *t*) for short (Fan et al. 2024a). It is the cornerstone of many important applications (Liang et al. 2023), such as question answering (Hao et al. 2017; Xu et al. 2022), information retrieval (Qin et al. 2023; Liu et al. 2018) and recommendation system (Zhang et al.

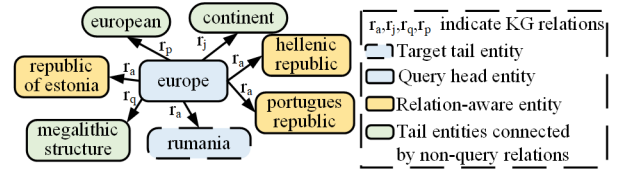


Figure 1: In this figure, the yellow bricks denote relation-aware entities which share the same head entity and relation in the testing triple. These entities share important concept information with the target entity (the dashed blue brick), thus can be used to guide for better link prediction.

2016; Chen et al. 2023). Due to the intricate relations between entities, manually constructed knowledge graphs often struggle to encompass knowledge comprehensively. As a consequence, knowledge graph completion, which uncovers missing links by reasoning over known facts, is an important task for enhancing the application effects in this field.

The existing methods in this field can be broadly divided into two categories, i.e., triple-based (Li et al. 2024; Wang et al. 2014) and text-based methods (Wang et al. 2022a; Shang et al. 2023). Algorithms in the second category, which utilize pre-trained language models to incorporate additional relations and entity descriptions for improved predictive accuracy, have garnered escalating interest in recent years. Specifically, the text embedding methods (Wang et al. 2022a; Yao, Mao, and Luo 2019), as typical representatives of text-based methods, are achieving promising performance with appealing efficiency by adopting a bi-encoder network structure and a contrastive learning manner. In these methods, the visible components of the triplets, either (*h*, *r*, *?*) or (*?*, *r*, *t*), together with the corresponding descriptions serve as queries, while the tail or head entities considered as the candidate values. Employing language models, this branch’s methods map queries and candidate values into a shared hidden space with structure identical but independent sub-networks. They perform link prediction by aligning the embeddings of valid triples and repelling those of the invalid ones. Our proposed method also falls into this group.

The performance of these methods is highly related to two vital factors. First, the representative capability of the pre-trained language model. To this aspect, thanks to the rapid

*These authors contributed equally.

†Corresponding author.

development of large language models (Devlin et al. 2019; OpenAI 2024), their stunning knowledge storage and context understanding ability have ensured the representative capability of the learned embeddings. Second, the precision of the query. While finding the appropriate value to match the query, it is easy to know that the more precise context information that the query provides, the easier the matching will be. Take the tail prediction task as an example. According to the setting of the task, the included descriptions of head entities and relations have largely specified these two components. However, to the tail entity, not even a word of the what the target entity looks like is provided. To solve the problem, the existing methods design various negative sampling algorithms to enable the learned network to better tell what does the valid tail entities not look like (Jiang, Drummond, and Cohn 2023; Li et al. 2023a; He et al. 2023; Je, Choi, and Oh 2023). Although good performance has been acquired, the existing methods neglect to provide context information of the tail entities.

In this paper, we find that the relation-aware entities can act as effective contexts of the target entities. In Figure. 1, we illustrate a sub-graph of the WN18RR dataset as an example. In this subgraph, *romania* is the target tail entity of the incomplete triplet (*europe*, *has part*, ?) in the testing set. The yellow entities represent the neighbors of the query head entity that share the same relation with the target tail entity. These entities (connected with the query head entity by the query relation) are denoted as relation-aware entities in our paper. Through this figure, we have two observations. 1) Relation-aware entities are more inclined to possess similar semantic concepts, introducing these entities as contexts could provide important guidance to the target entity prediction. For example, in this subgraph, when given the relation-aware entity *portugues republic*, it is more intuitive to anticipate the algorithm providing a European country as a response to the query (*europe*, *has part*, ?). These entities can help specify the query to make them to be more different from similar counterparts, like (*continent*, *has part*, ?). 2) Without introduction of the target entity, it is very likely to have multiple entities that is suitable for a query. For example, besides Rumania, Germany, and Spain are all suitable for the query. In this circumstance, a possible solution to improve the algorithm performance is to enhance the average matching degree of the whole relation-aware entity set.

Driven by the observations, we randomly sample entities from the relation-aware entity set as anchors to guide the query embedding learning. Specifically, in our method, we also adopt a bi-encoder network structure and a contrastive learning manner (Wang et al. 2022a). However, besides constructing classic queries which only contain information of head entities and relations, more context abundant anchor-enhanced queries are developed. These queries are formulated by appending anchor entities and their descriptions to the original query. With these queries, the network is trained by minimizing the InfoNCE loss (Chen et al. 2020) over both classic and anchor-enhanced query embeddings. In this setting, with the anchor information as extra context of the tail entity, the generated anchor-enhanced embedding is pulled to get closer to the neighborhood of the target en-

tity, thus being more discriminative over the link prediction task. Moreover, as the relation-aware entities are included in the query sequence, more relation compact representation will be acquired. The contributions of this paper are twofold:

- Through observation and experimental verification, we find a previously overlooked fact that the relation-aware entities can act as effective contexts of target entity prediction. Enhanced by this setting, the current leading methods can achieve significant performance improvements without requiring substantial modifications.
- We propose a relation-aware anchor enhanced knowledge graph completion (RAA-KGC) by complementing the queries with target entity prototypes. The proposed method is proven to generate discriminative and compact embeddings, thus achieves superior performance against the compared state-of-the-art algorithms.

Related Work

Knowledge Graph Completion

The KGC methodology has experienced rapid advancement in recent years. To the triple-based KGC methods, numerous strategies (i.e., translation-based (Bordes et al. 2013; Wang et al. 2014), semantic matching-based (Nickel, Tresp, and Kriegel 2011), neural network-based (Schlichtkrull et al. 2018; Li et al. 2023d; Fan et al. 2024b), logical-rule based (Sadeghian et al. 2019)) have been developed to construct suitable entity and relation embeddings while conducting proper assessment to the validation of the candidate triples. These models acquire the structural aspects of a knowledge graph, excluding the textual content of entities and relations. Comparatively, text-based methods suggest integrating textual knowledge of entities and relations into the model via Pre-trained Language Models (Zeng et al. 2024). The pioneer text-based method, DKRL (Xie et al. 2016), initially integrates textual descriptions into entity embeddings with a convolutional neural network. Subsequent studies, including KG-BERT (Yao, Mao, and Luo 2019) and KEPLER (Wang et al. 2021b), utilize a Pre-trained Language Model (PLM) for encoding text descriptions. Moreover, KGT5-context (Kochsiek et al. 2023) and KG-S2S (Chen et al. 2022a) model the link prediction task as a sequence-to-sequence generation problem and output the entity predictions by the decoder directly. Furthermore, LMKE (Wang et al. 2022b) introduces a contrastive learning framework that employs a PLM to derive entity and relation embeddings within the same space as word tokens, showcasing its efficacy in addressing the long-tail issue. Good performance have been achieved by the text-based algorithms, and the performance is highly correlated with the context information which is given to depict the target entities.

Context Information Adopted by KGC

Precise context information is important for the rationality and accuracy of the output of the language models, that is why prompt engineering is so important to neural language processing (Li et al. 2023c). To collect context information for the algorithms, a variety of methods have been

proposed. The early methods straightforwardly take the descriptions of query entities and relations as context. After that, algorithms are proposed to collect more context information within the neighborhood of the query triplet. Among them, KGT5-context (Kochsiek et al. 2023) takes the one-hop neighborhood of the query as extra context to the model. ConKGC (Shang et al. 2023) introduces a sampling strategy to extract weak positive samples within the relation-aware neighborhood of the query. BERTRL (Zha, Chen, and Yan 2022) collects possible reasoning paths in the query neighborhood for intrinsic supervision. Besides the mentioned algorithms, GenKGC (Xie et al. 2022) and KICGPT (Wei et al. 2023) go beyond the neighborhood of the query and take triples connected by the query relation as prediction demonstrations. Significant performance enhancement has been witnessed with the aforementioned context information extraction mechanism. However, these methods mainly focus on introducing more common knowledge, which is related to either the query entity or the query relation. In this paper, through our observation and experiments, we find that exploiting task-specific information besides the common knowledge information for the query is also important to the model performance. Integrating the proposed relation-aware anchor strategy with the existing mechanisms can further improve the algorithm performance. In the following section, we introduce the anchor generation procedure and the network structure of our method.

Method

Problem Statement

Given a knowledge graph (KG) $\mathcal{G} = \{(h, r, t) | h, t \in \mathcal{E}, r \in \mathcal{R}\}$, where \mathcal{E} and \mathcal{R} are the set of entity and relation of the KG, respectively. h and t are the head and tail entities, while r is the relation between them. The KGC tasks are usually divided into link prediction, triple classification, and relation prediction (Lin et al. 2015; Yao, Mao, and Luo 2019; Chen et al. 2022b). In this paper, we focus on the link prediction sub-task, which aims to predict the missing h in head entity prediction $(?, r, t)$ or t in tail entity prediction $(h, r, ?)$ for a relation fact triple (h, r, t) . The evaluation protocol is to rank all candidate entities instead of only giving one best result. Following the approach presented in SimKGC (Wang et al. 2022a), we use the inverse triples (t, r^{-1}, h) to simplify the task by focusing exclusively on tail entity prediction, where r^{-1} denotes the inverse relation of r . Besides, we also summarize the notations in the Appendix.

Anchor Generation

The success of RAA-KGC is attributed to leveraging information from relation-aware entities, specifically utilizing the neighbors of the query head entity that share the same relation with the target tail entity. The RAA-KGC framework is illustrated in Figure. 2. Based on KGs, we define anchor generation operation as follows.

Definition 1. Relation-Aware Anchor Generation

Given a knowledge graph $\mathcal{G} = \{(h, r, t) | h, t \in \mathcal{E}, r \in \mathcal{R}\}$ and the corresponding inverse knowledge graph $\mathcal{G}_{inv} = \{(t, r^{-1}, h) | \forall (h, r, t) \in \mathcal{G}\}$. When the query is $(h_a, r_a, ?)$,

the relation-aware entity set \mathcal{T} is the set of tail entities connected with the query head entity h_a and relation r_a on $\mathcal{G} \cup \mathcal{G}_{inv}$. The aim of anchor generation is to obtain anchor set \mathcal{T}_k , and the operation as follows.

$$\mathcal{T}_k = \text{random}(\mathcal{T}, k), k \leq K, K = 5 \quad (1)$$

where $\mathcal{T}_k = \{t_i | i = \{1, 2, \dots, k\}, t_i \in \mathcal{T}\}$, k is the random sampling size.

Given an input query (*europa, has part, ?*), the proposed method traverses all triples started from the query head entity h_a in the given KG and only keeps the entities satisfying Def. 1. After that, we construct the anchor-enhanced queries $I_{hrt_a}^a$ by equation (2) and the classic query I_{hr}^a by equation (3). We input these information into subsequent models.

$$I_{hrt_a}^a = [CLS]h_a : h_{adesc}[SEP]r_a[SEP] \\ t_i : t_{idesc}[SEP], i = \{1, 2, \dots, k\} \quad (2)$$

$$I_{hr}^a = [CLS]h_a : h_{adesc}[SEP]r_a[SEP] \quad (3)$$

$$I_t = [CLS]t : t_{desc}[SEP] \quad (4)$$

where h_a denotes query head entity, t_i is entity in anchor set \mathcal{T}_k , h_{adesc} and t_{idesc} are the descriptions of h_a and t_i , r_a is the relation between h_a and t_i , t and t_{desc} are candidate entities and their descriptions.

Anchor-Enhanced Embedding Generation

Following the current state-of-the-art KGC methods (Wang et al. 2022a), we use two encoders, $g_1(\cdot)$ and $g_2(\cdot)$, both initialized with the bert-base-uncased model but do not share parameters. The anchor-enhanced embedding is computed by Def.2. In addition, the I_{hr}^a is encoded by encoder $g_1(\cdot)$ to generate a relation-aware embedding e_{hr} . The second encoder $g_2(\cdot)$ is used to compute candidate embedding e_t for candidate entities sequence I_t .

Definition 2. Anchor-Enhanced Embedding Generation Given a knowledge graph $\mathcal{G} = \{(h, r, t) | h, t \in \mathcal{E}, r \in \mathcal{R}\}$ and the corresponding inverse knowledge graph $\mathcal{G}_{inv} = \{(t, r^{-1}, h) | \forall (h, r, t) \in \mathcal{G}\}$. $I_{hrt_a}^a$ is encoded by encoder $g_1(\cdot)$ as e_{hrt_a} . Then, the operation of generating anchor-enhanced embedding $e_{hrt_a}^{avg}$ as follows.

$$e_{hrt_a}^{avg} = \text{average}(e_{hrt_a}) \quad (5)$$

where e_{hrt_a} is anchor-enhanced queries embedding.

Training and Inference

We formulate a simple yet effective contrastive learning loss function based on the anchor-enhanced embedding, which uses cosine to measure the similarity between the target entity and the query (Huang et al. 2022). The function leverages the hidden semantics underlying the relation-aware entities, as shown in equations (6)-(8). The positive samples are the fact triples in KG, and the negative samples are the self-negative samples (SN) and in-batch negative samples (IBN) (Wang et al. 2022a). Moreover, different relations within the same batch will have different anchor sets,

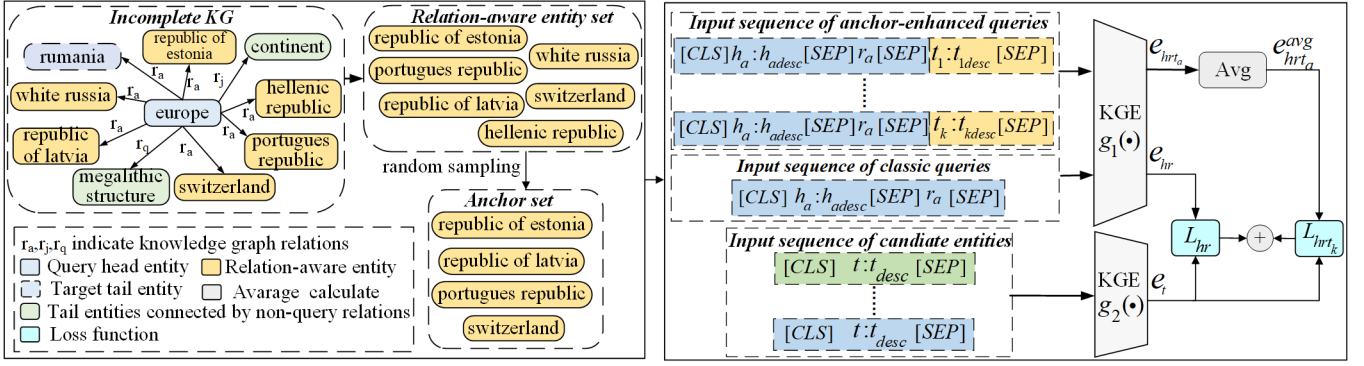


Figure 2: Framework illustration of the proposed RAA-KGC. In this figure, the yellow bricks denote relation-aware entities that share the same head entity and relation. The dashed blue brick denotes the target entity. The core idea of RAA-KGC is to generate a general example of what might the tail (or head) entity be like by using relation-aware entities, and then pull its embedding towards the neighborhood of the anchor.

leading to different anchor-enhanced embedding, even if the query head entity is the same. Therefore, we also incorporate in-batch-relation negative samples (IBRN). This IBRN method is similar to IBN. By minimizing L_{hrt_a} , the anchor-enhanced embedding $e_{hrt_a}^{avg}$ can be further rotated and scaled, thereby bringing the target entity embedding closer to $e_{hrt_a}^{avg}$ in the resulting semantic space. Based on this, when predicting $(h, r, ?)$, $e_{hrt_a}^{avg}$ and target entity embedding will get a higher score than others, so we can prioritize search results in these relation-aware entities. After the above operation, the generated anchor-enhanced embedding is pulled toward the neighborhood of the target entity. Then, it is positioned to the specific anchor through L_{hr} .

$$L_{cls} = \alpha \cdot L_{hrt_a} + L_{hr} \quad (6)$$

$$L_{hrt_a} = -\log \frac{e^{\phi(h,r,t_a)-\gamma)/\tau}}{e^{\phi(h,r,t_a)-\gamma)/\tau} + \sum_{t'_a \in \varphi} e^{\phi(h,r,t'_a)/\tau}} \quad (7)$$

$$L_{hr} = -\log \frac{e^{\phi(h,r,t)-\gamma)/\tau}}{e^{\phi(h,r,t)-\gamma)/\tau} + \sum_{t' \in \delta} e^{\phi(h,r,t')/\tau}} \quad (8)$$

where $\varphi = (\text{IBN} \cup \text{IBRN})$, $\delta = (\text{SN} \cup \text{IBN})$, γ is the additive margin, τ is a learnable temperature parameter, α represents the trade-off weight, $\phi(h, r, t_a)$ is cosine of $e_{hrt_a}^{avg}$ and e_t , $\phi(h, r, t)$ is cosine of e_{hr} and e_t .

For inference, we calculate the two cosine similarity, $\phi(h, r, t_a)$ and $\phi(h, r, t)$, and choose the sum of the highest score of each as the prediction:

$$\arg\max \phi(h, r, t_a) + \phi(h, r, t) \quad (9)$$

Experiments

This section is organized as follows. Firstly, we introduce the experimental settings. Then, we evaluate RAA-KGC's performance on three benchmark datasets. Finally, we conduct some experimental studies to investigate: (i) the importance

Dataset	train	valid	test
WN18RR	86,835	3034	3134
FB15k-237	272,115	17,535	20,466
Wikidata5M-Trans	20,614,279	5,163	5,163

Table 1: Statistics of the datasets used in this paper.

of anchor-enhanced embedding, (ii) relation-wised performance, (iii) the impact of sample size, (iv) the compatibility of the relation-aware anchor enhancement mechanism, (v) performance over unseen entities. (vi) case study.

Experimental Settings

Datasets We evaluate RAA-KGC on three commonly used datasets: WN18RR (Dettmers et al. 2018), FB15k-237 (Toutanova and Chen 2015), Wikidata5M-Trans (Wang et al. 2021b). The details of these three datasets are shown in Table 1. For WN18RR and FB15k-237 datasets, Teru et al. (Teru, Denis, and Hamilton 2020) extracted four versions of transductive and inductive datasets, respectively. For WN18RR and FB15k-237 datasets, we utilize the text descriptions provided by KG-BERT (Yao, Mao, and Luo 2019). The Wikidata5M-Trans dataset already contains descriptions for all entities and relations.

For the WN18RR dataset, SimKGC (Wang et al. 2022a) preprocessed entities as follows. If the entity is *reform_NN_1*, it is extracted as *reform*. A large number of similar entities, such as *reform_VB_1*, *reform_NN_3*, and *reform_VB_5*, exist with distinct meanings. These account for approximately 13.76% of the total entities. To preserve entity uniqueness, we do not apply this preprocessing.

Implementation Detail We employ the text-based model SimKGC (Wang et al. 2022a) as the baseline. The parameter settings for the KGE encoder are the same as shown in the original paper. We implement RAA-KGC based on the PyTorch library (Paszke et al. 2019). The batch size is 32. As for the specific hyperparameters used in our work, we

search the upper bound of the trade-off weight α for contrastive loss within the range $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. We assess RAA-KGC using four automated evaluation metrics: Mean Reciprocal Rank (MRR) and Hit@k(k=1,3,10).

Compared Baselines To demonstrate the superior performance of the proposed method, we compare it with twelve models. For triple-based methods, we report results of TransE (Bordes et al. 2013), ComplEx (Trouillon and Bouchard 2017), DistMult (Yang et al. 2015), ConvE (Dettmers et al. 2018), RGCN (Schlichtkrull et al. 2018), RotatE (Sun et al. 2019), and QuatE (Li et al. 2023b). For text-based methods, we include KG-BERT (Yao, Mao, and Luo 2019), MTL-KGC (Kim et al. 2020), StAR (Wang et al. 2021a), SimKGC (Wang et al. 2022a), and HaSa (Zhang, Zhang, and Molybog 2024). Note that except for the results of SimKGC, the results of the compared baselines are recorded from the original papers.

Link Prediction Performance

To evaluate the effectiveness of the RAA-KGC, we compare it against twelve SOTA methods over three KGC datasets, including triple-based and text-based methods. The main results are summarized in Table 2, from which several conclusions can be drawn. First, RAA-KGC outperforms prior works on the WN18RR and Wikidata5M-Trans datasets. Specifically, on the WN18RR dataset, it improves MRR by 4.43%, Hit@1 by 5.61%, Hit@3 by 3.44%, and Hit@10 by 2.36% compared to SimKGC. Second, we observe that RAA-KGC falls short on the FB15k-237 dataset, possibly because the topological structure of this dataset is considerably denser (Yang et al. 2024; Li et al. 2023a; Cao et al. 2021). Conversely, this kind of dataset is suitable for combining text and structure information for prediction (He et al. 2023), which we consider as a future research. Additionally, compared with QuatE in the triple-based methods, RAA-KGC improves MRR by 11.64%, Hit@1 by 7.04%, and Hit@10 by 19.64% on the WN18RR dataset. Similarly, compared with TransE, RAA-KGC increases MRR by 35.44%, Hit@1 by 46.34%, and Hit@10 by 22.81% on the WN18RR dataset. In the text-based methods, RAA-KGC achieves superior performance on all metrics of the WN18RR dataset and the Wikidata5M-Trans dataset. These results indicate entities similar to the target entity exist in the dataset, suggesting that leveraging these entities can enhance the model’s ability to learn entity characteristics.

Importance of Anchor-Enhanced Embedding

In order to further validate the importance of anchor-enhanced embedding, the following experiments are conducted: (1) NT: uses only relation-aware embedding, (2) NRT: uses only tail entities connected by non-query relations, (3) IET: uses only anchor-enhanced embedding, (4) IRT: uses anchor-enhanced embedding and relation-aware embedding. Table 3 shows these results. Besides, we select five groups of the tail entity sets from WN18RR dataset and embed them in both NT and IRT scenarios. The embeddings are visualized using t-SNE (Van der Maaten and Hinton 2008), as shown in Figure. 3.

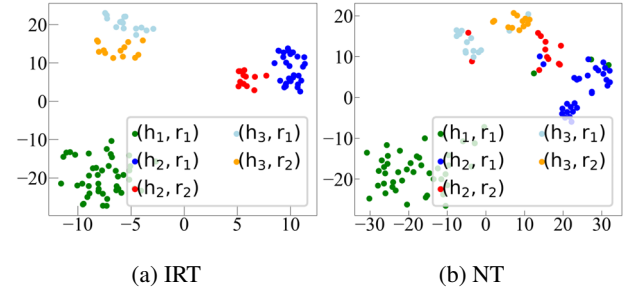


Figure 3: Visualization of entity embeddings with the different head entities and relations using t-SNE in NT and IRT scenarios. Points in same color indicate that their head entities and relations are the same. h_1, h_2, h_3 represent different query head entities, r_1 and r_2 indicate knowledge graph relations. The NMI of (a) and (b) are 0.75 and 0.62, respectively.

As shown in Table 3, the performance is optimal in the IRT scenario and minimal in the NRT scenario. These findings indicate that the relation-aware anchor plays an important role in the prediction process. Meanwhile, we can see the following points from Figure. 3: (1) The gap among tail entities, which are connected by the same query head entity and different relations, is smaller than the gap among tail entities corresponding to different query head entities. For example, the gap between (h_2, r_1) and (h_2, r_2) is smaller than that between (h_1, r_1) and (h_2, r_1) . Similar situations can be observed in the relations between other entities. (2) The comparison of IRT and NRT shows that the similarity between the neighbors of the query head entity and the target entity is higher than for others. Therefore, making full use of these entities allows the model to learn a general example of what might the tail (or head) entity be like and thus improves the model’s discrimination. (3) The superior performance of IRT compared to NT indicates that incorporating relation-aware entities offers the model additional supervision, thereby enhancing its performance.

Relation-Wised Performance

To explore which relation enhances the RAA-KGC performance from a fine-grained perspective, we compare the performance of the RAA-KGC and SimKGC for each relation on WN18RR dataset. The difference in Hit@1 between RAA-KGC and SimKGC is illustrated in Figure. 4. The details of MRR, Hit@1, Hit@3, and Hit@10 are provided in the Appendix.

As shown in Figure. 4, RAA-KGC outperforms SimKGC across most relation types, indicating the feasibility and effectiveness of our approach. Although it exhibits suboptimal performance on the *member_of_domain_region* relation, this relation types account for 1.06% of all triples. Overall, these fine-grained analysis results show that RAA-KGC achieves outstanding performance.

Impact of the Sample Size

We calculate the MRR and Hit@k(k=1,3,10) for different sample sizes on WN18RR datasets, as shown in Table 4.

Methods	WN18RR				FB15k-237				Wikidata5M-Trans			
	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10	MRR	Hit@1	Hit@3	Hit@10
Triple-based methods												
TransE	24.3	4.3	44.1	53.2	27.9	19.8	37.6	44.1	25.3	17.0	31.1	39.2
ComplEx	46.8	42.7	48.5	55.4	27.8	19.4	29.7	45.0	28.2	22.6	-	39.7
DistMult	44.4	41.2	47.0	50.4	28.1	19.9	30.1	44.6	33.9	27.1	-	45.9
ConvE	45.6	41.9	47.0	53.1	31.2	22.5	34.1	49.7	-	-	-	-
RGCN	42.7	38.2	44.6	51.0	24.8	15.3	25.8	41.4	-	-	-	-
RotatE	47.6	42.8	49.2	57.1	33.8	24.1	37.5	53.3	29.0	23.4	32.2	39.0
QuatE	48.1	43.6	-	56.4	31.1	22.1	-	49.5	-	-	-	-
Text-based methods												
KG-BERT	21.6	4.1	30.2	52.4	-	-	-	42.0	-	-	-	-
MTL-KGC	33.1	20.3	38.3	59.7	26.7	17.2	29.8	45.8	-	-	-	-
StAR	40.1	24.3	49.1	70.9	29.6	20.5	32.2	48.2	-	-	-	-
SimKGC	55.31	45.03	61.54	73.65	29.89	21.27	32.18	47.68	29.18	25.37	30.58	44.9
HaSa	53.8	44.4	58.8	71.3	30.4	22.0	32.5	48.3	-	-	-	-
RAA-KGC	59.74	50.64	64.98	76.01	31.46	22.38	34.06	49.26	34.15	29.24	34.47	47.82

Table 2: The results of WN18RR, FB15k-237, Wikidata5M-Trans datasets. Best performance is indicated by the bold face number in each column, and the underline means the best result compared to SimKGC in each column.

Metrics	NT	NRT	IET	IRT
MRR	55.31	50.99	51.66	59.74
Hit@1	45.03	40.03	41.05	50.64
Hit@3	61.54	57.21	60.47	64.98
Hit@10	73.65	71.77	74.13	76.01

Table 3: Results of adding different tail entity embeddings on the WN18RR dataset. The bold face number in each row indicates the best performance.

Metrics	0	1	2	3	4	5
MRR	55.31	55.69	55.75	58.08	59.74	59.56
Hit@1	45.03	45.36	46.19	48.63	50.64	50.48
Hit@3	61.54	62.76	63.99	63.78	64.98	64.93
Hit@10	73.65	73.87	74.46	75.19	76.01	75.62

Table 4: Impact of the sample size on WN18RR dataset. The bold face in each row indicates the best performance.

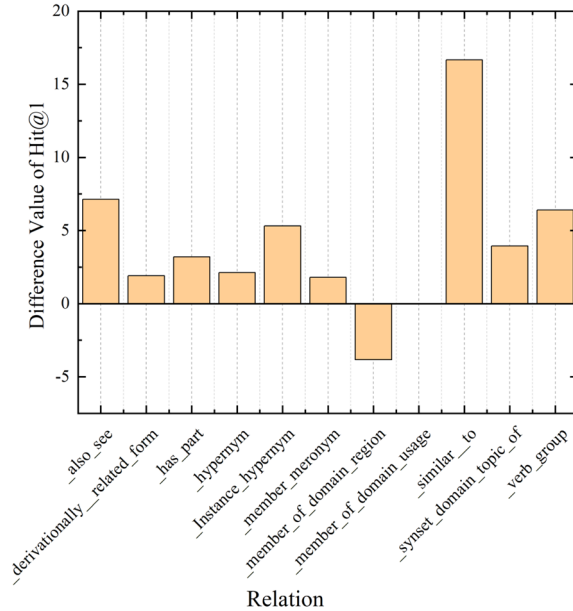


Figure 4: The difference value of Hit@1 between RAA-KGC and SimKGC on different relations in the WN18RR dataset.

By analyzing these results, the performance is poorest at a sample size of 0, with performance improving as the sample size increases. When the sample size is 4, Hit@10 reaches 76.01%, a 2.36% improvement over a sample size of 0 on WN18RR dataset. We attribute the improvement to the finding of anchor-enhanced embeddings, which capture more masked entities and contextual information for reconstructing the masked elements. However, as the sample size grows, the introduction of noise limits further performance gains. When the sample size is 5, Hit@10 is 75.62% on WN18RR datasets, which is 0.39% lower than the highest recorded values. The decline may be due to the presence of too many relation-aware entities, leading to a bias in the model’s cognition of tail entities.

Compatibility of the Relation-Aware Anchor Enhancement Mechanism

To assess the compatibility and plug-and-play attribute of our approach, Table 5 shows the performance of four baseline models w./w.o. our RAA framework on the WN18RR.v1 dataset (Teru, Denis, and Hamilton 2020). As for triple-based methods, we extend our method to two typical baselines, including TransE (Bordes et al. 2013) and ComplEx (Trouillon and Bouchard 2017). As for the text-based methods, we choose InsKGC (Jiang, Drummond, and Cohn 2023) and SimKGC (Wang et al. 2022a). Table 5 indicates that the RAA framework makes average boosts on all

Methods	MRR	Hit@1	Hit@3	Hit@10
TranE	16.54	1.3	25.86	52.50
TransE-RAA	17.49	2.1	26.96	54.55
InsKGC	59.55	49.61	65.52	77.03
InsKGC-RAA	62.92	54.31	67.94	78.84
ComplEx	53.06	48.98	55.96	59.25
ComplEx-RAA	55.33	50.96	57.13	62.07
SimKGC	56.23	46.63	61.83	73.83
SimKGC-RAA	63.21	54.47	69.04	77.51

Table 5: Experiments on compatibility of relation-aware anchor enhancement mechanism with other methods on WN18RR_v1 dataset. The bold numbers indicate the better performance.

Methods	WN18RR_v1		FB15k-237_v1	
	MRR	Hit@10	MRR	Hit@10
RAA-KGC	50.12	84.04	29.11	55.36
SimKGC	41.77	66.49	20.98	40.98
InsKGC	40.12	69.15	23.5	48.29

Table 6: The Hit@10 and MRR of WN18RR_v1 and FB15k-237_v1 under inductive scenario. The optimal values of each metric are marked in bold.

metrics for baselines. Concretely, with the help of our strategy, the performance of InsKGC improves by 3.37%, 4.7%, 2.42%, and 1.81% on MRR, Hit@1, Hit@3, and Hit@10, respectively. Additionally, SimKGC also shows improvements of 6.98%, 7.84%, 7.21%, and 3.68% on MRR, Hit@1, Hit@3, and Hit@10. Moreover, ComplEx and TransE also showed improvements on all metrics. These experimental results verify the compatibility of solutions against state-of-the-art algorithms and further suggest that the relation-aware entity information can enhance the discriminative ability of other KGC models.

Performance Over Unseen Entities

To further explore the generalization capability of RAA-KGC, we conduct experiments under an inductive link prediction setting. The texted datasets include WN18RR_v1, WN18RR_v2, WN18RR_v3, WN18RR_v4, FB15k-237_v1, FB15k-237_v2, FB15k-237_v3, and FB15k-237_v4. The eight datasets were extracted by Teru et al. (Teru, Denis, and Hamilton 2020). The compared algorithms include SimKGC and InsKGC (Jiang, Drummond, and Cohn 2023). Due to space limitations, only the Hit@10 and MRR of WN18RR_v1 and FB15k-237_v1 datasets are shown in Table 6, more results are presented in the Appendix. As shown in Table 6, RAA-KGC outperforms the other two methods on each metric. On the WN18RR_v1 dataset, RAA-KGC achieves 8.35% and 10% higher MRR than SimKGC and InsKGC, respectively. Similarly, on the FB15k-237 dataset, RAA-KGC is 8.13% and 5.61% higher than SimKGC and InsKGC on MRR. These results indicate that anchor-enhanced embedding provides more supervision information, enabling the network to better comprehend the semantic information of relations and entities. This allows RAA-

(almaty_NN_1, instance hypernym, urban_center_NN_1)		
Model	Top5 candidate entites	probabilities
SimKGC	national_capital_NN_1	0.78
	dtate_capital_NN_1	0.72
	capital_NN_3	0.65
	provincial_capital_NN_1	0.57
	urban_center_NN_1	0.56
RAA-KGC	urban_center_NN_1	0.82
	national_capital_NN_1	0.68
	port_NN_1	0.65
	capital_NN_3	0.57
	state_capital_NN_1	0.53

Table 7: The ranks predicted and probability by RAA-KGC and SimKGC. Bold indicates the target entity.

KGC to adapt quickly.

Case Study

We further intuitively show that the prediction results using *almaty_NN_1* as the query head entity and *instance hypernym* as the relation. The top-5 predicted entities and their corresponding probabilities, as ranked by RAA-KGC and SimKGC, are presented in Table 7. We notice that the RAA-KGC outperforms SimKGC, with the target entity appearing first. The target entity, *urban_center_NN_1*, represents a specific city, while the other predictions denote the broader category of capital. Note that we do not need to know the concrete label of the entity or the relation. What we need to know is that these entities belong to anchors referred to in Def. 1. It is because the relational edges in KGs naturally constrain these entities. Starting from the same query head entity and the same relation, we will reach the tail entities with similar semantics, no matter what the query head entity and relation are. Using such property of the RAA-KGC to construct the relation-aware anchor will be more stable since no other language models are required.

Conclusion

In this paper, we propose a relation-aware anchor enhanced knowledge graph completion method, termed RAA-KGC. This method first generates a general example of what might the tail (or head) entity be like, and then pulls its embedding towards the neighborhood of the anchor. Experiments reveal the effectiveness and compatibility of RAA-KGC. In the future, we aim to incorporate the structural and textual information of the subgraph to improve the performance of the KGE model.

References

- Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.
- Cao, Y.; Ji, X.; Lv, X.; Li, J.; Wen, Y.; and Zhang, H. 2021. Are Missing Links Predictable? An Inferential Benchmark for Knowledge Graph Completion. In *Proceedings of*

- the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 6855–6865.
- Chen, C.; Wang, Y.; Li, B.; and Lam, K.-Y. 2022a. Knowledge Is Flat: A Seq2Seq Generative Framework for Various Knowledge Graph Completion. In *Proceedings of the 29th International Conference on Computational Linguistics*, 4005–4017.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.
- Chen, X.; Zhang, N.; Li, L.; Deng, S.; Tan, C.; Xu, C.; Huang, F.; Si, L.; and Chen, H. 2022b. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, 904–915.
- Chen, Z.; Li, D.; Zhao, X.; Hu, B.; and Zhang, M. 2023. Temporal knowledge question answering via abstract reasoning induction. *arXiv preprint arXiv:2311.09149*.
- Dettmers, T.; Minervini, P.; Stenetorp, P.; and Riedel, S. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *NAACL*, 4171–4186.
- Fan, S.; Fan, G.; Nie, H.; Yao, Q.; Liu, Y.; Li, X.; and Wang, Z. 2024a. Flow to Candidate: Temporal Knowledge Graph Reasoning With Candidate-Oriented Relational Graph. *IEEE Transactions on Neural Networks and Learning Systems*.
- Fan, S.; Nie, H.; Wang, R.; Yao, Q.; Du, H.; Liu, Y.; and Wang, Z. 2024b. Relation-Entity Hybrid Learning Graph Model for Few-Shot Temporal Knowledge Graph Forecasting. In *International Conference on Database Systems for Advanced Applications*, 342–357. Springer.
- Hao, Y.; Zhang, Y.; Liu, K.; He, S.; Liu, Z.; Wu, H.; and Zhao, J. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 221–231.
- He, J.; Jia, L.; Wang, L.; Li, X.; and Xu, X. 2023. Mocosa: Momentum contrast for knowledge graph completion with structure-augmented pre-trained language models. *arXiv preprint arXiv:2308.08204*.
- Huang, S.; Tsang, I. W.; Xu, Z.; and Lv, J. 2022. Measuring Diversity in Graph Learning: A Unified Framework for Structured Multi-View Clustering. *IEEE Transactions on Knowledge and Data Engineering*, 34(12): 5869–5883.
- Je, S.-H.; Choi, W.; and Oh, K. 2023. Unifying Structure and Language Semantic for Efficient Contrastive Knowledge Graph Completion with Structured Entity Anchors. *arXiv preprint arXiv:2311.04250*.
- Jiang, F.; Drummond, T.; and Cohn, T. 2023. Don’t mess with mister-in-between: Improved negative search for knowledge graph completion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 1818–1832.
- Kim, B.; Hong, T.; Ko, Y.; and Seo, J. 2020. Multi-task learning for knowledge graph completion with pre-trained language models. In *Proceedings of the 28th international conference on computational linguistics*, 1737–1743.
- Kochsiek, A.; Saxena, A.; Nair, I.; and Gemulla, R. 2023. Friendly Neighbors: Contextualized Sequence-to-Sequence Link Prediction. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
- Li, H.; Wang, L.; Wei, Y.; Da Xu, R. Y.; and Wang, B. 2023a. KERMIT: Knowledge Graph Completion of Enhanced Relation Modeling with Inverse Transformation. *arXiv preprint arXiv:2309.14770*.
- Li, L.; Zhang, X.; Jin, Z.; Gao, C.; Zhu, R.; Liang, Y.; and Ma, Y. 2023b. Knowledge graph completion method based on quantum embedding and quaternion interaction enhancement. *Information Sciences*, 648: 119548.
- Li, M.; Wang, W.; Feng, F.; Cao, Y.; Zhang, J.; and Chua, T.-S. 2023c. Robust Prompt Optimization for Large Language Models Against Distribution Shifts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 1539–1554. Association for Computational Linguistics.
- Li, Q.; Guo, S.; Wu, J.; Li, J.; Sheng, J.; Peng, H.; and Wang, L. 2023d. Event extraction by associating event types and argument roles. *IEEE Transactions on Big Data*.
- Li, Q.; Li, J.; Wu, J.; Peng, X.; Ji, C.; Peng, H.; Wang, L.; and Philip, S. Y. 2024. Triplet-aware graph neural networks for factorized multi-modal knowledge graph entity alignment. *Neural Networks*, 106479.
- Liang, K.; Liu, Y.; Zhou, S.; Tu, W.; Wen, Y.; Yang, X.; Dong, X.; and Liu, X. 2023. Knowledge graph contrastive learning based on relation-symmetrical structure. *IEEE Transactions on Knowledge and Data Engineering*, 36(1): 226–238.
- Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*.
- Liu, Z.; Xiong, C.; Sun, M.; and Liu, Z. 2018. Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval. *arXiv preprint arXiv:1805.07591*.
- Nickel, M.; Tresp, V.; and Kriegel, H.-P. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning, ICML’11*, 809–816.
- OpenAI. 2024. GPT-4 Technical Report. *arXiv:2303.08774*.

- Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.
- Qin, L.; Li, Z.; Yu, Q.; Wang, L.; and Che, W. 2023. Towards complex scenarios: Building end-to-end task-oriented dialogue system across multiple knowledge bases. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 13483–13491.
- Sadeghian, A.; Armandpour, M.; Ding, P.; and Wang, D. Z. 2019. DRUM: end-to-end differentiable rule mining on knowledge graphs. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 15347–15357.
- Schlichtkrull, M.; Kipf, T. N.; Bloem, P.; Van Den Berg, R.; Titov, I.; and Welling, M. 2018. Modeling relational data with graph convolutional networks. In *The semantic web: 15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, proceedings 15*, 593–607. Springer.
- Shang, B.; Zhao, Y.; Wang, D.; and Liu, J. 2023. Relation-aware multi-positive contrastive knowledge graph completion with embedding dimension scaling. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 878–888.
- Sun, Z.; Deng, Z.-H.; Nie, J.-Y.; and Tang, J. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *International Conference on Learning Representations*.
- Teru, K.; Denis, E.; and Hamilton, W. 2020. Inductive relation prediction by subgraph reasoning. In *International Conference on Machine Learning*, 9448–9457. PMLR.
- Toutanova, K.; and Chen, D. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, 57–66.
- Trouillon, T. P.; and Bouchard, G. M. 2017. Complex embeddings for simple link prediction. US Patent App. 15/156,849.
- Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).
- Wang, B.; Shen, T.; Long, G.; Zhou, T.; Wang, Y.; and Chang, Y. 2021a. Structure-augmented text representation learning for efficient knowledge graph completion. In *Proceedings of the Web Conference 2021*, 1737–1748.
- Wang, L.; Zhao, W.; Wei, Z.; and Liu, J. 2022a. SimKGC: Simple Contrastive Knowledge Graph Completion with Pre-trained Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4281–4294. Dublin, Ireland: Association for Computational Linguistics.
- Wang, X.; Gao, T.; Zhu, Z.; Zhang, Z.; Liu, Z.; Li, J.; and Tang, J. 2021b. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9: 176–194.
- Wang, X.; He, Q.; Liang, J.; and Xiao, Y. 2022b. Language Models as Knowledge Embeddings. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, 2291–2297. International Joint Conferences on Artificial Intelligence Organization.
- Wang, Z.; Zhang, J.; Feng, J.; and Chen, Z. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*.
- Wei, Y.; Huang, Q.; Zhang, Y.; and Kwok, J. 2023. KICGPT: Large Language Model with Knowledge in Context for Knowledge Graph Completion. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 8667–8683.
- Xie, R.; Liu, Z.; Jia, J.; Luan, H.; and Sun, M. 2016. Representation learning of knowledge graphs with entity descriptions. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2659–2665.
- Xie, X.; Zhang, N.; Li, Z.; Deng, S.; Chen, H.; Xiong, F.; Chen, M.; and Chen, H. 2022. From discrimination to generation: Knowledge graph completion with generative transformer. In *Companion Proceedings of the Web Conference 2022*, 162–165.
- Xu, X.; Qin, L.; Chen, K.; Wu, G.; Li, L.; and Che, W. 2022. Text is no more enough! a benchmark for profile-based spoken language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11575–11585.
- Yang, B.; Yih, S. W.-t.; He, X.; Gao, J.; and Deng, L. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *Proceedings of the International Conference on Learning Representations (ICLR) 2015*.
- Yang, R.; Zhu, J.; Man, J.; Fang, L.; and Zhou, Y. 2024. Enhancing text-based knowledge graph completion with zero-shot large language models: A focus on semantic enhancement. *Knowledge-Based Systems*, 300: 112155.
- Yao, L.; Mao, C.; and Luo, Y. 2019. KG-BERT: BERT for knowledge graph completion. *arXiv preprint arXiv:1909.03193*.
- Zeng, Z.; Chen, Y.; Yu, S.; Yan, Y.; Liu, Z.; Wang, S.; Han, X.; Liu, Z.; and Sun, M. 2024. KBAda: Efficient Self Adaptation on Specific Knowledge Bases. *arXiv preprint arXiv:2411.14790*.
- Zha, H.; Chen, Z.; and Yan, X. 2022. Inductive relation prediction by BERT. In *Proceedings of the AAAI conference on artificial intelligence*, 5923–5931.
- Zhang, F.; Yuan, N. J.; Lian, D.; Xie, X.; and Ma, W.-Y. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 353–362.
- Zhang, H.; Zhang, J.; and Molybog, I. 2024. HaSa: Hardness and Structure-Aware Contrastive Knowledge Graph Embedding. In *Proceedings of the ACM on Web Conference 2024*, 2116–2127.