Journal of Computers Vol. 37 No. 4, August 2026, pp. 1-12
https://doi.org/10.63367/19911599202608370400 1

Research on Entity Alignment Models for Multimodal Knowledge Graphs

Chunxia Lan[1*], Jianxin Li[2], and Jiajun Li[2]

[1] School of Teacher Training, Ningbo Polytechnic University,
Ningbo, 315800, China
274556104@qq.com

[2] School of Electronic Information, Dongguan Polytechnic,
Dongguan, 523808, China
279149042@qq.com, 3345340890@qq.com

Received 10 May 2025; Revised 15 July 2025; Accepted 4 August 2025

Abstract: The entity alignment task in Multi modal Knowledge Graph (MKG) faces challenges such as strong modal heterogeneity, large semantic gap, and inflexible fusion strategies. Existing methods generally rely on static weighted fusion and large amounts of annotated data, making it difficult to cope with complex multimodal environments. This article proposes a novel multimodal entity alignment model- TextFusionEA, which implements structural embedding based on GraphSAGE, uses MobileNetV3 to obtain visual features, introduces an adaptive fusion mechanism of semantic, structural, and visual features, and designs a semi supervised iterative learning strategy to extend the seed subset to alleviate the problem of data scarcity. Experiments on five datasets have shown that TextFusionEA performs well in MRR Hits@1. It significantly outperforms existing mainstream methods in terms of metrics, verifying its alignment performance and robustness in complex multimodal scenarios.

Keywords: multimodal knowledge graph, entity alignment, GraphSAGE, adaptive fusion, semi-supervised learning

1 Research Background and Significance

In the context of the rapid growth of multi-source heterogeneous information today, knowledge graph, as an important supporting technology for semantic understanding and reasoning, is widely used in fields such as search recommendation and question answering systems [1]. However, existing knowledge graphs are mostly independently constructed by different organizations, with a large number of semantically overlapping but formally heterogeneous entities [2]. How to achieve cross graph entity alignment has become the core task of knowledge fusion. Especially in multimodal knowledge graphs, entities are not only described by text, but also contain information such as images and audio. Entity alignment tasks further face the following challenges:

The large semantic gap between different modalities makes it difficult to represent information uniformly;

The modal fusion strategy lacks adaptability and often relies on static weights;

Lack of training data, high cost of manual annotation, and constraints on alignment performance.

Currently, most research focuses on single modality or only uses simple modality fusion methods, lacking deep integration of structural information and multiple semantic features, and having weak utilization capabilities for a small number of seed entities. Therefore, this article proposes the following research questions around the above challenges:

How to design an adaptive modal feature fusion mechanism that effectively combines image, text, and structural information?

How to expand high-quality entity pairs through semi supervised learning under the condition of scarce seed entities?

This study takes this as a starting point to construct a multimodal entity alignment model that considers flexibility, generalization, and scalability, aiming to improve the alignment performance and practicality in real complex graphs.

* Corresponding Author

With the advancement of artificial intelligence, knowledge graphs, leveraging their powerful semantic understanding capabilities, have provided new approaches for artificial intelligence and natural language processing. Knowledge graphs represent knowledge using graph structures, organizing and associating entities, relationships, and attributes as nodes and edges. They are commonly expressed in the "subject-predicate-object" triple form, such as the knowledge graph for "Liu Dehua" being constructed through triples like (Liu Dehua, blood type, AB type). As shown in Fig. 1.

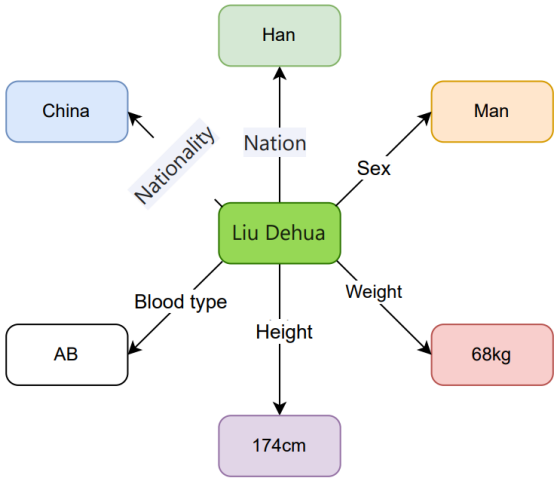

Fig. 1. Example of the knowledge graph for the "Andy Lau" entry

As AI advances rapidly, the applications of knowledge graphs continue to expand. However, different knowledge graphs suffer from heterogeneity and data redundancy issues, and the information from a single knowledge graph is limited, making it necessary to integrate multiple knowledge graphs. Entity alignment is a crucial step in knowledge graph integration. For instance, Table 1 and Table 2 demonstrate an entity alignment example. Suppose there are two knowledge graphs, KG1 and KG2. The entity information contained in KG1 is as follows:

Table 1. Entity information KG1

Entity ID	Entity name	Attribute 1	Attribute 2
E1	Apple Inc	Technology company	Headquartered in the United States
E2	iPhone 14	Mobile phone model	Equipped with high-definition camera

The entity information contained in KG2 is shown in Table 2.

Table 2. Entity information KG2

Entity ID	Entity name	Attribute A	Attribute B
E3	Apple Inc.	High tech enterprises	A well-known American company
E4	iPhone 14 Pro	Intelligent mobile phone	Equipped with advanced camera function

In this example, "Apple Inc. (E1)" in KG1 and "Apple Inc. (E3)" in KG2 can be aligned as entities because they both represent the same real-world entity, Apple Inc. Through entity alignment, related entities in the two knowledge graphs can be linked, achieving knowledge fusion and integration.

Today's data exhibits multimodal characteristics, including text, images, audio, and video. In a multimodal environment, the same or related entities may not be consistent across different modal knowledge graphs. Therefore, researching multimodal knowledge graph entity alignment models is of great significance, as it contributes to the advancement of knowledge graph technology and artificial intelligence.

2 Related Work

2.1 Research Status at Home and Abroad

Entity alignment in knowledge graphs has evolved from traditional rule-based methods to representation learning approaches that better address the challenges of large-scale, multi-modal, and semantically rich graphs. Graph neural networks (GNNs), including GCN [3] and GAT [4], effectively capture graph structures for entity embedding but face issues such as computational overhead and poor interpretability in multi-modal settings [5]. Translation-based models like TransE, TransH, and TransR offer scalable and efficient alignment by embedding entities and relations into low-dimensional vector spaces [3, 6], though they remain sensitive to hyperparameters and lack adaptability to dynamic scenarios.

To improve semantic alignment, researchers have incorporated local semantic features using CNNs, RNNs, and attention mechanisms [7]. These enhancements improve expressiveness but still face limitations in semantic disambiguation and integration with graph structure. In multi-modal knowledge graphs, the heterogeneity of textual and visual data introduces further complexity. Approaches using attention-based fusion [7] and generative adversarial networks (GANs) [8] show promise, yet challenges in cross-modal consistency and generalization persist.

Fusion strategies have matured from static concatenation to dynamic attention-based and Transformer-driven architectures. For instance, Cong et al. [9] proposed modality-aware attention weighting, and Wang et al. [10] introduced cross-modal alignment modules, while Transformer-based designs enable higher-order interaction modeling [11]. Despite these improvements, scalability and robustness remain issues. In parallel, semi-supervised learning has become increasingly important. Techniques such as BootEA's iterative seed expansion [12], NAEA's neighborhood attention [13], and the noise-resistant mechanisms in EVA [14] and RNM [15] help reduce annotation costs but still risk negative transfer. In response, our work proposes an adaptive dual-channel embedding and robust seed expansion strategy to further improve multi-modal alignment performance.

2.2 Current Challenges and Proposed Solutions in Multimodal Entity Alignment

Existing research on entity alignment in multimodal knowledge graphs faces three key limitations: (1) inadequate exploration of local semantic information beyond structural features, (2) inflexible fixed-weight fusion approaches that fail to capture inter-modal correlations, and (3) insufficient training data coupled with high annotation costs for complex entity relationships. To address these challenges, this study proposes a comprehensive framework featuring: (1) deep analysis of multimodal feature representations and their intrinsic relationships, with experimental validation of image-assisted alignment effectiveness; (2) an adaptive feature fusion mechanism for dynamic weight optimization; and (3) a semi-supervised iterative alignment method designed to reduce data dependency while improving model performance.

3 Entity Alignment Model for Multimodal Knowledge Graph

3.1 Overall Architecture of the Model

At present, knowledge graph entity alignment faces three main problems: most of the work focuses on textual information, neglecting other modalities such as images; Multimodal fusion typically uses fixed weights, which limits the performance of the model; The lack of labeled data in the graph representation learning method leads to performance degradation. Therefore, this article constructs a multimodal knowledge graph entity alignment model. The overall framework of the model is shown in Fig. 2. This model uses GraphSAGE to model the structure of knowledge graphs, with two multimodal knowledge graphs DMT1 as inputs. DMT2 embeds them into the structure to obtain a structural similarity matrix. Then, treating relationships and attributes as bag of words features, modeling both and inputting them into a feedforward neural network to obtain relationship embeddings. Obtain entity name features through GloVe vector averaging and input them into a feedforward neural network to obtain the corresponding embeddings. For visual information, this article uses a pre trained visual model MobileNetV3 to obtain visual features, and inputs the visual features into a feedforward neural

network to obtain visual information embedding. Then, the relationship embedding, attribute embedding, entity name embedding, and visual embedding are fused to generate semantic embeddings, and the fused semantic similarity matrix is obtained by calculating cosine similarity. By using an adaptive weighting strategy to integrate the structure and semantic matrix, an entity similarity matrix is obtained. Finally, iteratively filter reliable entity pairs to expand the seed set until the model converges, supporting unsupervised alignment.

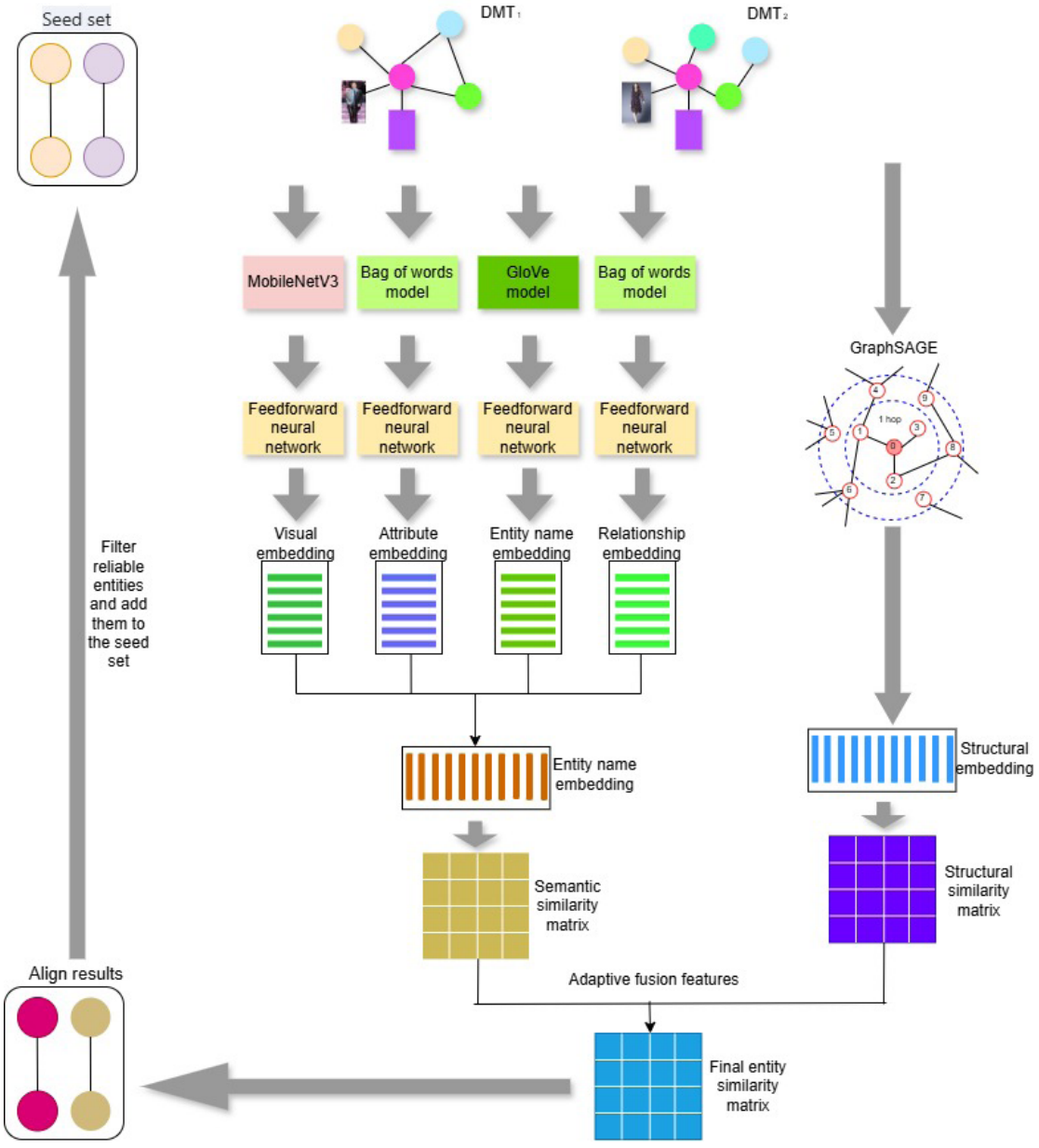

Fig. 2. Overall framework diagram of the model

3.2 Structural Feature Embedding

GraphSAGE stands for Graph Sampling and Aggregation Network, which is a neural network model used for processing large-scale graph structured data. This article uses GraphSAGE to model the structural information of knowledge graphs. In entity alignment tasks, GraphSAGE performs multi-level neighbor random sampling on

target entities (such as structured neighbors across modalities such as text and images) in multimodal knowledge graphs, and dynamically fuses semantic information of the target entity and its multimodal neighbors using multi-layer feature aggregation (such as mean or pool aggregation) to generate robust embedding representations, as shown in Formula 1.

$$H^{(l+1)} = \sigma(W^{(l)} \bullet CONCAT(H^{(l)}, \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)})) \cdot \quad (1)$$

In equation 1, $H^{(l)}$ and $H^{(l+1)}$ represent the node characteristic matrix of layer l and layer l+1 respectively, σ is the relu activation function, $\tilde{D}$ represents the degree matrix of the node, D represents the adjacency matrix, and I is the identity matrix. $W^{(l)}$ Is a randomly initialized weight matrix. In this study, the output of the last graphsage layer is embedded as the structure of the knowledge map.

After getting the vector representation of each entity through graphsage model, for the distance between entities v_i and v_j, this study uses Manhattan distance to calculate the similarity of structure embedding:

$$d(v_i, v_j) = \sum_{k=1}^{d} \left| h_{v_i}^{(k)} - h_{v_j}^{(k)} \right| . \quad (2)$$

In equation 2, v_i is the entity in the knowledge map $DMT1$, and v_j is the entity in the knowledge map $DMT2$, h_{vj} representing the structure vector of entity v_i. The smaller the value of $d(v_i, v_j)$, the higher the structural similarity of the two entities v_i and v_j. Then, the entity structure vectors generated by the graphsage models are mapped into a common vector space by using the seed set for subsequent entity alignment operations.

3.3 Embedding Features of Relationships, Attributes, and Entity Names

Although structural features are key information for entity alignment in knowledge graphs, relying solely on them cannot fully understand the rich semantics of entities, which may lead to incorrect alignment. Therefore, this article utilizes semantic information from relationships, attributes, and entity names to assist in entity alignment. Relationships and attributes are considered as bag of words features, and feedforward neural networks are used to calculate relationship and attribute embeddings. The calculation is as follows:

$$r_e^i = t_e p_e^i + h_e . \quad (3)$$

$$r_a^i = t_a p_a^i + h_a . \quad (4)$$

Among them, the R_i relationship embedding r_a^i as an entity, the R_i bag of words relationship feature vector p_e^i as an entity, t_e and h_e as learnable parameters. p_a^i r_a^i As R_i attribute embeddings R_i of entities, as bag of words attribute feature vectors of entities, t_a and h_a are learnable parameters. From this, two feature matrices can also be obtained: the relational feature matrix R_E and the attribute feature matrix R_A .

The initial feature vectors of entity names are obtained through the GloVe model and then input into a feedforward neural network to generate entity name embeddings. This study also draws on the research findings of Mao et al. [16], utilizing binary tuples as auxiliary features for entity name embeddings. The calculation formula is as follows:

$$r_n^i = t_n p_n^i + h_n . \quad (5)$$

In equation 5, r_n^i as the R_i embedding of entity names, p_n^i as the R_i initial feature vector of entity names, t_n and h_n are learnable parameters. Next, we can further derive the entity name feature matrix R_N .

3.4 Visual Feature Embedding

The method for obtaining visual feature embeddings is similar to the aforementioned approach for acquiring text-based information feature embeddings. Specifically, this study employs the MobileNetV3 visual model to obtain visual feature vectors, which are then fed into a feedforward neural network to derive visual feature embeddings:

$$r_v^i = t_v p_v^i + h_v . \tag{6}$$

In equation 6, r_v^i as the R_i visual feature information embedding of the entity, p_v^i is the R_i visual feature of the entity, t_v and h_v is the learnable parameter. Finally, the visual feature matrix can be obtained R_V .

3.5 Adaptive Feature Fusion

Existing multi-modal entity alignment methods are mostly based on representation-level fusion, which has certain limitations. Inspired by the research of Zeng [17] and others, this paper adopts a fusion method based on the similarity matrix level. It aggregates relations, entity names, attributes, and visual embeddings to obtain semantic embeddings and a semantic similarity matrix. This matrix is then adaptively fused with the structural similarity matrix to derive the final entity similarity matrix.

This study aggregates the relationship embedding R_E, attribute embedding R_A, entity name embedding, R_N and visual embedding obtained earlier R_V to obtain semantic embedding R_S . Its calculation method is as follows:

$$R_S = \underset{m \in M}{\oplus} \left[\frac{e^{wm}}{\sum_{i \in M} e^{wi}} R_m \right]. \tag{7}$$

Among them $M = \{E, A, N, V\}$, w_m is the m weight that can be trained by the modality. Before performing weighted concatenation, it is necessary to normalize each input feature matrix. Then, the semantic similarity matrix is calculated in the same way M_S, and the sum is fused with fixed weights M_S to obtain the final similarity matrix M.

$$M^{ij} = M_H^{ij} \times weight_H + M_S^{ij} \times weight_S . \tag{8}$$

In equation 8, $weight_H$ and $weight_S$ represent the weights of structural and semantic information, respectively, and M^{ij} represent the j final similarity scores of entities i and entities.

In practical applications, there are many limitations to assigning fixed weights to structural and semantic information. Therefore, this study designed an adaptive feature fusion method:

$$weight_H = \frac{a}{1 + b \times R^{-c(\deg + nei)}} . \tag{9}$$

$$weight_S = 1 - weight_H . \tag{10}$$

In equation 9, a, b, and c are hyperparameters, deg represents the degree of entity, *nei* represents the degree of association between the entity and the seed entity, and *nei* is calculated as follows:

$$nei = n_1 \times q_1 + \lg(n_2 \times q_2) . \tag{11}$$

Among them, $n_1 n_2$ represents the number of entities, q_1 and q_2 are hyperparameters.

3.6 Semi Supervised Entity Alignment Strategy

The size of the seed set directly affects the effectiveness of entity alignment. This study adopts an iterative alignment strategy to expand the seed set. Firstly, set a threshold and add entity pairs with similarity exceeding this threshold to the seed set, which are considered reliable entity pairs. Then, these newly generated entities are used for iterative training until the required number is reached. During the iteration process, it is inevitable to encounter unreliable entity pairs, which may affect subsequent training. In order to mitigate the impact of unreliable entity pairs, this study proposes the following strategies to address the aforementioned issues. Choose the bidirectional entities that are most similar to each other. This entity has the characteristic of directional asymmetry. Therefore, in order to minimize the occurrence of such errors, this study only considers body piles that satisfy the following two equations as reliable solid pairs. The official representatives are as follows:

$$R_i = \arg\max(\underset{R_m \in DMT1}{lik}(R_m, R_j)) . \tag{12}$$

$$R_j = \arg\max(\underset{R_n \in DMT2}{lik}(R_i, R_n)) . \tag{13}$$

Among them, *lik*(,) representing the similarity between entities, this study R_j considers entity pairs that satisfy the above two equations as reliable entity pairs.

Setting a similarity threshold is a strategy that involves defining similarity values to filter reliable entity pairs. This involves setting similarity values to perform secondary filtering on entity pairs. Only when an entity pair meets the most similar criteria and their similarity exceeds the set threshold, will these reliable entity pairs be added to the seed set.

The candidate entity pair strategy includes first adding entity pairs that satisfy the most similar entity policy and similarity threshold policy to the candidate entity pair set. If these candidate entity pairs still meet the criteria of being the most similar to each other and meeting the similarity threshold after several rounds of proposals, they are considered reliable entity pairs and added to the seed set.

4 Experiment and Analysis

4.1 Data Sets and Evaluation Indicators

This study selected five datasets as experimental datasets, including three cross language datasets DBP15KZH-EN (Chinese English), DBP15KJA-EN (Japnese English), and DBP15KFR-EN (French English), as well as two multimodal knowledge graph datasets FB15K-DB15K and FB15K-YAGO15K. This study utilized the images, entities, relations, attributes, relation triples, attribute triples, and entity pair counts from the above five datasets. Table 3 presents the dataset information.

Table 3. Dataset information

Data set		Image	Entity	Relationship	Attribute	Relationship triplet	Attribute triplet	Entity logarithm
DBP15K ZH-EN	ZH	15893	19398	1698	8230	69987	247432	14900
	EN	14201	19842	1322	7010	96783	342908	
DBP15K FR-EN	FR	14214	19763	980	4546	11023	265987	15000
	EN	13698	19602	1268	6231	116840	360123	
DBP15K JA-EN	JA	12802	19803	1287	5802	76435	243563	15000
	EN	13232	19453	1123	6032	93456	324562	
FB15K-DB15K	FB15K	13202	14879	1368	129	604531	29456	12960
	DB15K	12042	12653	298	221	87345	48032	
FB15K-YAGO15K	FB15K	13021	14239	1340	109	586754	29031	11868
	YAGO15K	12032	15132	42	11	122365	23543	

For those entities without images, this study assigns random vectors to them. In the DBP15K dataset, this study used 30% aligned entities as seed sets, and in the multimodal knowledge graph dataset, 20%, 60%, and 90% aligned entities were used as seed sets, respectively.

The evaluation metrics used in this study are Hits@k and Mean Reciprocal Rank (MRR). Hits@k is primarily used to measure the frequency or proportion of relevant items appearing in the top k results. In this study, it represents the proportion of correctly aligned entities among the top K candidate entities (with k set to 1 and 10). The value of Hits@k indicates the effectiveness of entity alignment.

4.2 Experimental Environment Setup

The experimental environment and parameter settings for this study are shown in Table 4.

Table 4. Experimental environment

Experimental environment	Configuration setting
Operating system	Ubuntu 20.04 LTS
CPU	Intel Core i9-13900K @ 3.5 GHz
Memory	256GB
GPU	NVIDIA GeForce RTX 4090Ti
Python	Python 3.11
Deep learning framework	Torch1.13.0
Related dependencies	NumPy 1.26.0, Matplotlib 3.8.0, Scikit-learn 1.3.0

4.3 Experimental Results and Analysis

This study selected nine entity alignment modes, namely GMNN [18], AliNet [19], CEAFF [20], BootEA [12], MMREA [21], NAEA [13], EVA [14], RNM [15], PoE [22], HMEA, and MME [23], as comparative models. And divide the above models into three groups for experimentation: non iterative entity alignment model, iterative entity alignment model, and multimodal entity alignment model.

Table 5 and Table 6 show the experimental results of the text model and the comparison model. Bold numbers in the table indicate the best model performance, underlined numbers indicate the second best model performance, and "/" indicates indicators not mentioned. The evaluation indicators are expressed as percentages.

Table 5. Experimental results of text model, non iterative and iterative entity alignment model on DBP15K dataset

Model	FR-EN			ZH-EN			JA-EN		
	H@1	H@10	MRR	H@1	H@10	MRR	H@1	H@10	MRR
Experimental results of text model and non iterative entity alignment model on DBP15K dataset									
GMNN	60.2	86.7	70.3	46.7	69.6	48.9	45.7	72.8	58.9
AliNet	58.6	84.6	65.4	52.8	82.8	62.6	56.8	83.9	64.7
EVA	73.2	93.8	77.8	74.4	<u>92.4</u>	<u>79.3</u>	72.3	<u>93.1</u>	79.6
CEAFF	**93.7**	<u>95.4</u>	**95.2**	<u>73.9</u>	86.3	78.6	<u>81.2</u>	92.5	<u>84.8</u>
No iteration Text model	<u>89.8</u>	**96.6**	<u>90.3</u>	**79.8**	**93.9**	**85.8**	**83.8**	**94.8**	**87.3**
Experimental results of text model and iterative entity alignment model on DBP15K dataset									
MMREA	66.3	86.5	/	65.1	85.3	/	62.5	84.6	/
NAEA	62.7	87.7	76.3	65.6	86.4	72.4	64.2	87.4	71.3
EVA	67.4	95.3	85.2	76.2	91.2	81.3	76.8	91.1	81.5
RNM	**93.6**	**97.9**	<u>93.3</u>	<u>84.2</u>	<u>91.7</u>	<u>87.2</u>	<u>87.4</u>	<u>94.2</u>	<u>87.6</u>
Iteration text model	<u>92.7</u>	<u>96.4</u>	**95.9**	**86.6**	**94.7**	**89.6**	**88.6**	**95.5**	**91.3**

Table 5 presents the experimental data of the text model and entity alignment model on the DBP15K dataset. From the results, it can be seen that both the non iterative text model and the iterative text model achieved the best experimental results. Experimental data strongly demonstrates the effectiveness of text fusion of visual

information for entity alignment. Meanwhile, it can also be seen from the table that the experimental data of the iterative text alignment model is superior to that of the non iterative text alignment model, indicating that using an iterative strategy is more effective for entity alignment.

Table 6. Experimental results of text model and multimodal entity alignment model on cross graph datasets

Seed collection	Model	FB15K-DB15K			FB15K-YAGO15K		
		H@1	H@10	MRR	H@1	H@10	MRR
20%	PoE	13.7	25.5	17.3	11.2	22.4	14.9
	HMEA	12.5	36.6	/	10.7	31.4	/
	MMEA	26.5	54.6	35.3	23.6	48.3	30.7
	EVA	14.8	32.3	20.3	9.3	27.9	15.8
	Text model	**39.2**	**65.3**	**47.8**	**34.2**	**61.6**	**44.6**
60%	PoE	46.3	65.7	53.2	34.9	53.7	40.9
	HMEA	26.4	57.7	/	27.3	59.2	/
	MMEA	41.9	71.3	51.7	41.6	65.1	48.2
	EVA	22.2	47.4	30.8	24.8	47.3	32.6
	Text model	**50.8**	**71.7**	**57.9**	**45.6**	**71.4**	**55.9**
90%	PoE	67.3	82.3	72.5	57.5	75.3	63.6
	HMEA	42.1	78.4	/	43.2	80.6	/
	MMEA	58.7	**86.7**	68.8	59.9	**83.9**	**68.5**
	EVA	37.2	58.3	44.7	39.6	61.5	48.1
	Text model	**69.3**	83.9	**74.2**	**61.7**	78.6	67.2

Table 6 presents the experimental data of text models and multi-modal entity alignment models under different proportions of seed sets on two cross-graph datasets. From the experimental results, text models generally achieved the best performance across different evaluation metrics. Taking the MRR metric as an example, when the seed set was 20%, the performance improvement of the text model was the most significant, with the MRR metric improving by 10.2%-12.1% compared to the best performing contrast model. However, as the proportion of the seed set increased, the magnitude of performance improvement for the text model gradually decreased. Through experiments, it was found that the number of training seeds for entity alignment is not necessarily the more the better. Similarly, by comparing the experimental data of the H@1 and H@10 metrics, the same conclusion can be drawn.

4.4 Analysis of the Impact of Seed set Quantity

In order to investigate the impact of seed set size on text models, this study conducted experiments using an unsupervised entity alignment model with pre aligned seeds obtained from visual information.

Fig. 3. shows the experimental results on the FR-EN, JA-EN, and ZH-EN datasets of DBP15K.

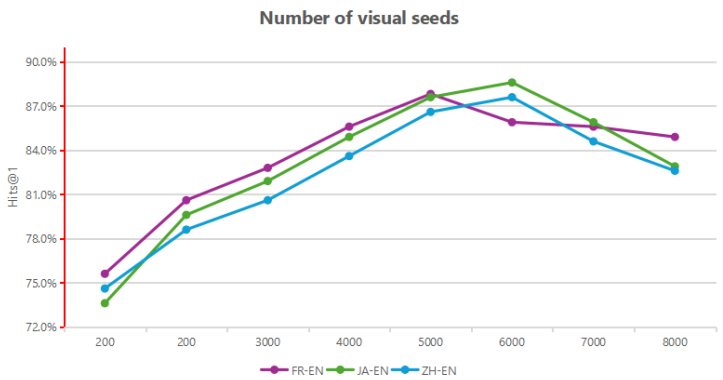

Fig. 3. Experimental results on FR-EN, JA-EN, and ZH-EN datasets of DBP15K

In Fig. 3, the purple, green, and blue lines represent the experimental results on FR-EN, JA-EN, and ZH-EN respectively. It can be seen from the figure that the three broken lines are on the rise, indicating that the performance of the model is gradually improved with the increase of the number of visual seeds. However, there are inflection points in all three lines, indicating that when the number of visual seeds reaches a certain amount, the model at this time Hits@1. When the maximum value is reached, the model is the best. The inflection points for the FR-EN, JA-EN, and ZH-EN datasets are 5,000, 6,000, and 6,000 respectively. The reason for the above phenomenon is that the seeds introduce noise, which makes the performance of the model begin to decline. When the number of seeds is 200, the Hits@1 values all exceed 74%, indicating that the text model can achieve good alignment performance using a small number of seed entities, thereby validating the effectiveness of the text model.

4.5 Ablation Experiment

To verify the effectiveness of each module of the text model, ablation experiments were designed on the DBP15K (JA-EN) dataset. Table 7 shows the relevant experimental data.

Table 7. Ablation experiment of text model on DBP15K (JA-EN)

Model	H@1	H@10	MRR
Entity name	80.8	95.6	86.7
Relationship	91.3	96.9	93.4
Attribute	89.5	93.6	91.8
Structure	54.1	81.3	63.2
Visual information	89.1	95.2	90.3
Iterative strategy	84.3	93.6	87.8
Adaptive feature fusion	89.3	96.2	91.5
Complete text model	94.2	98.2	96.8

As shown in Table 7 taking the Hits@1 metric as an example, the adaptive feature fusion strategy improves performance by 4.9%, which validates its effectiveness. The iterative strategy achieves a performance improvement of 9.9%, which is relatively smaller compared to improvements in other studies. This is due to the fact that visual and other supervisory information have already been well-aligned, thereby diminishing the performance enhancement brought by the iterative strategy. Regarding the impact of different modalities of information, the model's performance drops the most, by 50.1%, when structural information is removed. This indicates that structural information is critical for entity alignment in text models. The aforementioned results also demonstrate that different modalities of information have varying degrees of influence on entity alignment.

5 Summary and Prospect

This study conducted in-depth research on the entity alignment model of multimodal knowledge graphs and proposed an architecture for entity alignment in multimodal knowledge graphs. This model utilizes entity names, relationships, attributes, structures, and visual information in the knowledge graph to improve the effectiveness of entity alignment. The effectiveness of the model and its modules has been validated through multiple datasets. During the research process, several challenges were encountered. Firstly, the focus is only on the image information in multimodal data, without exploring audio or video modalities. Future research could consider integrating more types of modalities to further improve alignment accuracy. Secondly, noise in text image information can lead to low-quality visual data, affecting alignment performance. Future work can explore noise reduction technology to improve the quality of visual information.

Acknowledgement

This research was funded by the Key project of Zhejiang Provincial Education Science Planning 2025 Key Project (2025SB094), Zhejiang Province Higher Vocational Education "14th Five-Year Plan" Second Batch Teaching Reform Project (jg20240072), the Special for Key Fields of colleges and Universities in Guangdong Province under Grant (2024ZDZX1091), Dongguan Science and Technology of Social Development Program (20231800903592), Special Fund for Science and Technology Innovation Strategy of Guangdong Province (Special Fund for Climbing Plan)(pdjh2024b663).

References

[1] Y.W. Zhang, Q. Zhou, W. Chen, L. Zhao, Entity Alignment on Multi-modal Knowledge Graph, Journal of Chinese Computer Systems 45(5)(2024)1257-1263.
https://doi.org/10.20009/j.cnki.21-1106/TP.2022-0643

[2] W. Liu, H. Xu, W.M. Li, A Multimodal Knowledge Graph Entity Alignment Method, Journal of Applied Sciences 42(6) (2024) 1040-1051.
https://doi.org/10.3969/j.issn.0255-8297.2024.06.012

[3] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge graph embedding: A survey of approaches and applications, IEEE Transactions on Knowledge and Data Engineering 29(12)(2017) 2724-2743.
https://doi.org/10.1109/TKDE.2017.2754499

[4] Z.C. Wang, Q.S. Lv, X.H. Lan, Y. Zhang, Cross-lingual Knowledge Graph Alignment via Graph Convolutional Networks, in: Proc. the 2018 Conference on Empirical Methods in Natural Language Processing, 2018.
https://doi.org/10.18653/v1/D18-1032

[5] Z.H. Wu, S.R. Pan, F.W. Chen, G.D. Long, C.Q. Zhang, P.S. Yu, A comprehensive survey on graph neural networks, IEEE Transactions on Neural Networks and Learning Systems 32(1)(2021) 4-24.
https://doi.org/10.1109/TNNLS.2020.2978386

[6] W. Xiong, H. Chen, H.Y. Su, QuatCNNEx: A Knowledge Graph Embedding Model for Multiple Relation Mode, Chinese Journal of Computers 48(1)(2025) 124-135.
https://doi.org/10.11897/SP.J.1016.2025.00124

[7] D. Zhang, P. Liang, J.G. Gu, Multi-modal Knowledge Graph Link Prediction Based on Dual Fusion and Graph Attention Networks, Computer Technology and Development 34(7)(2024) 123-130.
https://doi.org/10.20165/j.cnki.ISSN1673-629X.2024.0089

[8] G. Feng, R.T. Zheng, T.X. Liu, Y.R. Yang, J.Z. Lin, T. Zhong, R.C. Huang, F. Xiang, W.C. Li, Multimodal entity alignment based on dual-generator shared-adversarial network, Application Research of Computers 42(6)(2025) 1632-1640.
https://doi.org/10.19734/j.issn.1001-3695.2024.11.0467

[9] S. Cong, G.B. Su, L. Liu, H.L. Wang, Review of Knowledge Graph Entity Alignment Research: from Traditional Methods to Cutting-Edge Technologies, Computer Engineering and Applications 62(1)(2026) 47-67.
https://doi.org/10.3778/j.issn.1002-8331.2412-0384

[10] H. Wang, L.J. Song, F. Du, Chinese Cross-modal Entity Alignment Method Based on Multi-modal Knowledge Graph, Computer Engineering 49(12)(2023) 88-95.
https://doi.org/10.19678/j.issn.1000-3428.0066938

[11] H. Li, X. Chen, T. Qu, H. Zhao, Multimodal text information extraction algorithm based on knowledge graph, Journal of Jilin University 56(6)(2026) 1687-1694.
https://doi.org/10.13229/j.cnki.jdxbgxb.20250103

[12] Z.Q. Sun, W. Hu, Q.H. Zhang, Y.Z. Qu, Bootstrapping entity alignment with knowledge graph embedding, in: Proc. the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018.
https://doi.org/10.24963/ijcai.2018/611

[13] Q.N. Zhu, X.F. Zhou, J. Wu, J.L. Tan, L. Guo, Neighborhood-Aware Attentional Representation for Multilingual Knowledge Graphs, in: Proc. the Twenty-Eighth International Joint Conference on Artificial Intelligence, 2019.
https://doi.org/10.24963/ijcai.2019/269

[14] F.Y. Liu, M.H. Chen, D. Roth, N. Collier, Visual pivoting for (unsupervised) entity alignment, in: Proc. The AAAI Conference on Artificial Intelligence, 2021.
https://doi.org/10.1609/aaai.v35i5.16550

[15] Y. Zhu, H. Liu, Z. Wu, Y. Du, Relation-Aware Neighborhood Matching Model for Entity Alignment, in: Proc. the AAAI Conference on Artificial Intelligence, 2021.
https://doi.org/10.1609/aaai.v35i5.16606

[16] X. Mao, W.T. Wang, Y.B. Wu, M. Lan, From alignment to assignment: Frustratingly simple unsupervised entity alignment, in: Proc. the 2021 Conference on Empirical Methods in Natural Language Processing, 2021.

https://doi.org/10.18653/v1/2021.emnlp-main.226
[17] W.X. Zeng, X. Zhao, X.Y. Li, J.Y. Tang, W. Wang, On entity alignment at scale, The VLDB Journal 31(5)(2022) 1009-1033.
https://doi.org/10.1007/s00778-021-00703-3
[18] K. Xu, L.W. Wang, M. Yu, Y.S. Feng, Y. Song, Z.G. Wang, D. Yu, Cross-lingual knowledge graph alignment via graph matching neural network, in: Proc. the 57th Annual Meeting of the Association for Computational Linguistics, 2019.
https://doi.org/10.18653/v1/P19-1304
[19] Z.Q. Sun, C.M. Wang, W. Hu, M.H. Chen, J. Dai, W. Zhang, Y.Z. Qu, Knowledge graph alignment network with gated multi-hop neighborhood aggregation, in: Proc. the 34th AAAI Conference on Artificial Intelligence, 2020.
https://doi.org/10.1609/aaai.v34i01.5354
[20] W.X. Zeng, X. Zhao, J.Y. Tang, X.M. Liu, P. Groth, Reinforcement Learning-based Collective Entity Alignment with Adaptive Features, ACM Transactions on Information Systems 39(3)(2021) 1-31.
https://doi.org/10.1145/3446428
[21] X.F. Shi, Y.H. Xiao, Modeling multi-mapping relations for precise cross-lingual entity alignment, in: Proc. The 2019 Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing, 2019.
https://doi.org/10.18653/v1/D19-1075
[22] Y. Liu, H. Li, A. Garcia-Duran, M. Niepert, D. Onoro-Rubio, D.S. Rosenblum, MMKG: Multi-Modal Knowledge Graphs, in: Proc. 2019 European Semantic Web Conference, 2019.
https://doi.org/10.1007/978-3-030-21348-0_30
[23] C.L. Ren, Z.K. Yu, C. Guan, L.Z. Jing, Multi-model entity alignment model based on adaptive fusion technology, Application Research of Computers 42(1)(2025) 100-105.
https://doi.org/10.19734/j.issn.1001-3695.2024.05.0187