

Article

Integration of Semantic and Topological Structural Similarity Comparison for Entity Alignment without Pre-Training

Yao Liu ¹  and Ye Liu ^{2,*} ¹ School of Computer Science and Technology, Beihang University, Beijing 100191, China; ly0708@buaa.edu.cn² School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing 100191, China

* Correspondence: lylyzyq1314@buaa.edu.cn

Abstract: Entity alignment (EA) is a critical task in integrating diverse knowledge graph (KG) data and plays a central role in data-driven AI applications. Traditional EA approaches rely on entity embeddings, but their effectiveness is limited by scarce KG input data and representation learning techniques. Large language models have shown promise, but face challenges such as high hardware requirements, large model sizes and computational inefficiency, which limit their applicability. To overcome these limitations, we propose an entity-alignment model that compares the similarity between entities by capturing both semantic and topological information to enable the alignment of entities with high similarity. First, we analyze descriptive information to quantify semantic similarity, including individual features such as types and attributes. Then, for topological analysis, we introduce four conditions based on graph connectivity and structural patterns to determine subgraph similarity within three hops of the entity's neighborhood, thereby improving accuracy. Finally, we integrate semantic and topological similarity using a weighted approach that considers dataset features. Our model requires no pre-training and is designed to be compact and generalizable to different datasets. Experimental results on four standard EA datasets validate the effectiveness of our proposed model.

Keywords: entity alignment; knowledge graph; description information; topological structure



Citation: Liu, Y.; Liu, Y. Integration of Semantic and Topological Structural Similarity Comparison for Entity Alignment without Pre-Training. *Electronics* **2024**, *13*, 2036. <https://doi.org/10.3390/electronics13112036>

Academic Editor: Arkaitz Zubiaga

Received: 1 May 2024

Revised: 17 May 2024

Accepted: 21 May 2024

Published: 23 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The entity-alignment (EA) [1,2] task involves matching and aligning entities from various data [3] sources or Knowledge graphs (KGs) [4–6]. Its objective is to establish associations between entities that share the same semantic meaning. This enables accurate correspondence when performing operations such as data integration [7], querying [8] and analysis [9] across different data sources. EA plays a crucial role in natural language processing, information extraction [10], intelligent question-answering systems [11] and social network analysis [12].

In the field of EA, knowledge-representation methods are primarily employed to obtain the vector information of entities for alignment purposes. Existing EA methods can be broadly categorized into three groups: translation-based methods, graph neural network (GNN)-based methods and other approaches. Translation-based methods, such as MTransE [13] and BootEA [14], use the TransE [15] framework to transform individual elements in the KG into semantically rich vector representations. This facilitates tasks such as semantic relevance reasoning, entity relationship reasoning and knowledge graph applications. GNN-based methods [16], represented by models such as GCN-Align [17], RDGCN [18] and Dual-AMN [19], generate entity embeddings by aggregating neighborhood information using graph neural networks. These models effectively capture structural information and learn entity embeddings. Other GNN-based methods, such as TEA-GNN [20], TREA [21] and STEA [22], further improve performance by incorporating temporal features. Other approaches, including Fualign [23], Simple-HHEA [24] and BERT-INT [25], address the challenge of heterogeneity in KGs by exploiting side information.

However, these three types of EA methods have significant limitations. They rely heavily on the performance of the Knowledge Representation Learning (KRL) model and the quality of the datasets used. Additionally, when different data sources have distinct mapping spaces within the KRL model, performing similarity comparisons becomes challenging.

Recently, large-scale language models (LLMs) [26] have demonstrated exceptional performance in various natural language processing tasks, including XLNet [27], GPT [28], Transformer-XL [29] and RoBERTa [30]. These transformer-based models take advantage of rich contextual information from large corpora and have proven to be highly effective in improving entity-related tasks in KGs. As a result, EA models based on LLMs have emerged, such as LLMEA [7], ChatEA [31] and BERT-INT [25], which effectively leverage the semantic and ancillary information provided by LLMs to perform entity-alignment (EA) tasks. However, it is important to recognize that while LLMs trained on large-scale textual data have rich linguistic patterns and statistical information, they may have limitations when dealing with ambiguity and polysemy. In such cases, LLMs tend to favor the most frequent or common meanings, potentially overlooking other possible interpretations in the given context. This limitation [32] becomes particularly significant in entity-alignment tasks, where a precise understanding of contextual information is crucial for accurately defining an entity. Alignment accuracy depends directly on this understanding, and therefore the limitations of LLMs can affect overall alignment performance.

In this paper, we present our entity-alignment model, which integrates semantic and structural information without the need for pre-training. The descriptive information of an entity, e.g. Beijing, the capital and largest city of China, with geographic coordinates (39°54' N, 116°23' E). Beijing holds significant political, cultural, and educational importance as the center of China.... Utilizing this descriptive information, we compare entities for semantic similarity. However, entities in the graph not only have semantic attributes but also exhibit structural characteristics, including attributes and features of neighboring nodes. To capture these characteristics, we propose a structure-based similarity-comparison module. This module takes into account the features of neighboring nodes and attributes of edges to determine whether entities in different graphs share similar structural features, employing conditional comparisons. Finally, the contributions of the two modules are effectively fused by employing a weighted sum approach, which takes into consideration the characteristics of the dataset. We evaluate the effectiveness of our approach on several EA datasets, including the traditional DBP15K (EN-FR) and DBP-WIKI datasets, as well as the more challenging and practical ICEWS-WIKI and ICEWS-YAGO datasets. These datasets exhibit a high degree of KG heterogeneity and complexity in capturing correlations between KGs. Through extensive experiments, we demonstrate that our method outperforms existing state-of-the-art EA models. Our contributions can be summarized as follows:

- We propose a novel EA model that employs both semantic and structural similarity comparisons. The entities are enhanced through the integration of semantic and structural information, thereby achieving highly accurate alignment. Furthermore, weighting factors are introduced to effectively balance the contributions of the two models, ensuring optimal alignment across different dataset features.
- We conduct EA task-based experiments on four datasets, and the results of these experiments demonstrate the effectiveness of our model.

The remainder of the paper is structured as follows: Section 2 provides a comprehensive explanation of our model. Section 3 presents the experimental setup. Section 4 shows the results and thoroughly analyzes these results. Section 5 concludes the paper.

2. Methods

This section describes the details of our approach. The overall framework is shown in Figure 1. Our model consists of three primary modules: semantic-based similarity comparison, topology-based similarity comparison and fusion of semantic and structural similarity comparison.

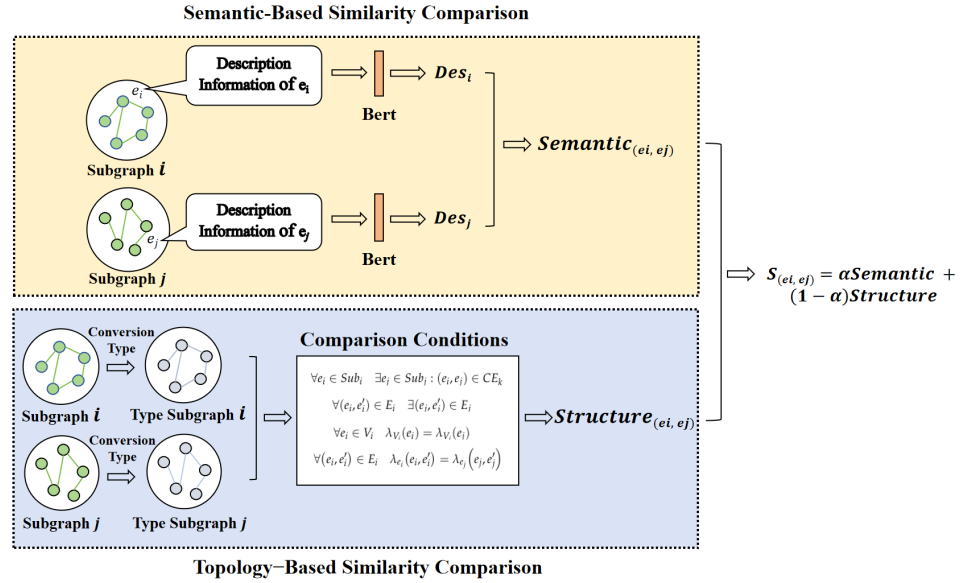


Figure 1. An overview of our model. The framework of our model consists of three components: Semantic-based Similarity Comparison, Structural-based Similarity Comparison and Fusion of Semantic and Structural Similarity Comparison. The Semantic-based Similarity Comparison module employs Bert to quantify the descriptive information of entities from different KGs and calculates their similarity using the cosine function. The Topological-based Similarity Comparison module determines the similarity of two subgraphs based on four specific conditions. Finally, the results of the two comparisons are integrated.

2.1. Preliminaries

Knowledge graph (KG). We define a KG as $\mathcal{G} = \{\mathcal{E}, \mathcal{R}\}$, where \mathcal{E} represents the set of entities and \mathcal{R} represents the set of relations. In a KG, a fact or edge is represented as a triple (h, r, t) , where $h \in \mathcal{E}$ denotes the head entity, $t \in \mathcal{E}$ denotes the tail entity and $r \in \mathcal{R}$ denotes the relation between them. The embedding vectors for h, r and t are denoted by \mathbf{h}, \mathbf{r} , and \mathbf{t} , respectively, using bold characters.

Entity alignment (EA). EA is a crucial task in KG research. Given two KGs, $\mathcal{G}_1 = (\mathcal{E}_1, \mathcal{R}_1)$ and $\mathcal{G}_2 = (\mathcal{E}_2, \mathcal{R}_2)$, the goal is to determine the identical entity set $S = \{(e_i, e_j) | e_i \in \mathcal{E}_1, e_j \in \mathcal{E}_2\}$. In this set, each pair (e_i, e_j) represents the same real-world entity but exists in different KGs.

2.2. Semantic-Based Similarity Comparison

To semantically measure whether two entities are similar, we quantify the descriptive information of the entities using Bert, and then test the quantified similarity. Specifically, consider the entities $e_i \in G_i, e_j \in G_j, e_i$ descriptive information labeled as Des_{e_i} and e_j descriptive information labeled as Des_{e_j} . The Bert employs the Transformer architecture, which is capable of processing longer text sequences and enabling better understanding of contextual information.

The vectors \mathbf{Des}_{e_i} and \mathbf{Des}_{e_j} are obtained after quantization by Bert. To test the similarity between the two vectors, we employ cosine similarity, which offers several advantages: simplicity, efficiency and independence of length and dimension in measuring vector similarity. These advantages align well with the experimental needs of our study. The cosine-based vector similarity calculation is:

$$Semantic_{(e_i, e_j)} = \frac{\mathbf{Des}_{e_i} \cdot \mathbf{Des}_{e_j}}{\|\mathbf{Des}_{e_i}\| \|\mathbf{Des}_{e_j}\|} \tag{1}$$

where $Semantic_{(e_i, e_j)}$ represents the similarity between e_i and e_j, \cdot denotes the dot product of vectors and $\|\mathbf{Des}_{e_i}\|$ and $\|\mathbf{Des}_{e_j}\|$ denote the norms of the vectors \mathbf{Des}_{e_i} and \mathbf{Des}_{e_j} respectively.

2.3. Topology-Based Similarity Comparison

An entity in a KG contains not only its semantic information, but also additional information such as topology. Entities in different KGs often share similarities in their links and relations. For example, an entity classified as person may have relations such as co-worker, colleague and more. These relations also connect to entity types that share similarities. Consequently, even when dealing with different KGs, the relations and associated entity types tend to have similarities. Leveraging this insight allows us to achieve entity alignment across different KGs.

We employ K-means to group entities based on their quantized descriptive information, allowing us to label entities in the same cluster as belonging to the same type. For example, consider the triple (Beijing, belongs_to, China), where both Beijing and China belong to the geography category. After clustering, these entities are assigned to cluster CE_0 . Similarly, the relationship 'belongs_to' is assigned to cluster CR_1 . Therefore, the type of the triple can be represented as (CE_0, CR_1, CE_0) .

We consider the entities and relations within three hops of the neighbor of the target entity to form a subgraph. The similarity between these subgraphs is then determined by evaluating their isomorphism.

Given two entities $e_i \in G_i$ and $e_j \in G_j$ with corresponding subgraphs $Sub_i = (V_i, R_i)$ and $Sub_j = (V_j, R_j)$ respectively, where V is the set of entities and R is the set of relations. The isomorphism between these two subgraphs can be determined based on the following conditions.

Condition 1:

$$\forall e_i \in Sub_i \quad \exists e_j \in Sub_j : (e_i, e_j) \in CE_k \quad (2)$$

where CE_k is the k th cluster. It is required that both e_i and e_j are of the same type.

Condition 2:

$$\forall (e_i, e'_i) \in E_i \quad \exists (e_j, e'_j) \in E_j \quad (3)$$

where e' denotes the neighborhood of node e . (e, e') denotes the edge connected between e and e' . The formula indicates that the edge present in the subgraph Sub_i are also present in Sub_j .

Condition 3:

$$\forall e_i \in V_i \quad \forall e_j \in V_j \quad \lambda_{V_i}(e_i) = \lambda_{V_j}(e_j) \quad (4)$$

where V_i denotes the set of nodes in Sub_i and V_j denotes the set of nodes in Sub_j . $\lambda_{V_i}(e_i)$ denotes the type of neighboring nodes of e_i , where the nodes all belong to V_i . This condition indicates that the types of all connected nodes of $e_i \in Sub_i$ and the types of the nodes connected to $e_j \in Sub_j$ should be similar.

Condition 4:

$$\forall (e_i, e'_i) \in E_i \quad \forall (e_j, e'_j) \in E_j \quad \lambda_{E_i}(e_i, e'_i) = \lambda_{E_j}(e_j, e'_j) \quad (5)$$

where E_i denotes the set of edges in Sub_i , while E_j denotes the set of edges in Sub_j . $\lambda_{E_i}(e_i, e'_i)$ denotes the type of neighboring edge of e_i , where the edges all belong to E_i . This condition indicates that the types of all connected edges of $e_i \in Sub_i$ and the types of all edges connected to $e_j \in Sub_j$ should be similar.

The two graphs can be considered isomorphic if and only if all of the above conditions are satisfied simultaneously.

2.4. Fusion of Semantic and Structural Similarity Comparison

In real-world scenarios, certain datasets are incomplete, resulting in a sparse structure. Therefore we introduce a parameter $\alpha \in [0, 1]$ to balance the semantic and structural information. The formula is defined as follows:

$$S_{(e_i, e_j)} = \alpha \times (Semantic_{(e_i, e_j)}) + (1 - \alpha) \times (Structure_{(e_i, e_j)}), \alpha \in [0, 1] \quad (6)$$

where $Structure_{(e_i, e_j)}$ denotes the structural similarity, which is calculated on the basis of the degree of coverage of the four conditions. For example, the third condition, $e_i \in Sub_i$ has approximately 80% of the number of nodes of the same type as $e_j \in Sub_j$. Therefore, the structural similarity can be defined by the following formula:

$$Structure_{(e_i, e_j)} = \alpha_1 \times (\#Edge) + \alpha_2 \times (\#Type_{node}) + \alpha_3 \times (\#Type_{edge})$$

$$\alpha_1 + \alpha_2 + \alpha_3 = 1 \quad (7)$$

where $\#Edge$ denotes the number of identical edges (i.e., Condition 2), $\#Type_{node}$ denotes the number of neighbors of the same type (i.e., Condition 3) and $\#Type_{edge}$ denotes the number of neighboring edges of the same type (i.e., Condition 4). The symbols α_1 , α_2 and α_3 denote the weights of the three conditions. These parameters are defined in a predetermined manner, and when the similarity between entities e_i and e_j , denoted as $S_{(e_i, e_j)}$, exceeds a certain threshold (e.g., $S_{(e_i, e_j)} > 0.9$), it indicates a high degree of similarity. Based on this similarity measure, entities e_i and e_j can be determined to be the same entity.

3. Experiment Setting

We conduct comparison experiments on the EA task using four standard datasets. Furthermore, we perform ablation studies to evaluate the specific contributions of two modules in our model. The effect of hyper-parameters on model performance is evaluated. The experiments are conducted on an Ubuntu 18.04.4 environment with 62 GB of RAM and a 60 GB GPU.

3.1. Datasets

We conduct experiments on four entity-alignment datasets. The statistics of these selected datasets are summarized in Table 1.

Table 1. The detailed statistics of the datasets. Temporal denotes whether the dataset contains temporal information.

Dataset		#Entities	#Relations	#Facts	Density	#Anchors	Temporal
DBP15K(EN-FR)	EN	15,000	193	96,318	6.421	15,000	No
	FR	15,000	166	80,112	5.341		No
DBP-WIKI	DBP	100,000	413	293,990	2.940	10,000	No
	WIKI	100,000	261	251,708	2.517		No
ICEWS-WIKI	ICEWS	11,047	272	3,527,881	319.352	5058	Yes
	WIKI	15,896	226	198,257	12.472		Yes
ICEWS-YAGO	ICEWS	26,863	272	4,192,555	156.072	18,824	Yes
	YAGO	22,734	41	107,118	4.712		Yes

DBP15K(EN-FR) and DBP-WIK [33] are two simple EA datasets, which share a similar structure for their KG pairs, with an equivalent number of entities. Furthermore, the structural features, such as the number of facts and density, of these two datasets closely align. ICEWS-WIKI and ICEWS-YAGO [24] are two complex EA datasets. Here, the KG pairs exhibit significant heterogeneity, differing not only in the number of entities but also in structural features. Notably, the quantity of anchors does not equal the number of entities. Consequently, aligning these complex datasets poses greater challenges.

To ensure consistency, we standardized the average length of descriptive information for entities across all datasets to 800 words. In cases where entities had insufficient descriptive information, we utilize ChatGPT [28] to supplement and enhance the available information.

3.2. Baselines

We select a set of state-of-the-art EA methods. These methods encompass a wide range of input features. Among the selected methods, we included translation-based approaches such as MTransE [13] and BootEA [14]. Additionally, we incorporated GNN-based techniques such as GCN-Align [17], RDGCN [18] and Dual-AMN [19]. Each baseline model

has unique characteristics and high performance that make them suitable for evaluating the effectiveness of our proposed method.

3.3. Hyper-Parameters

To determine the optimal hyperparameters for each dataset, we perform a grid search. The optimal experimental setup includes the following hyperparameters:

For DBP15K(EN-FR): $\alpha = 0.6$, $\alpha_1 = 0.4$, $\alpha_2 = 0.3$, $\alpha_3 = 0.3$. For DBP-WIKI: $\alpha = 0.6$, $\alpha_1 = 0.3$, $\alpha_2 = 0.4$, $\alpha_3 = 0.3$. For ICEWS-WIKI: $\alpha = 0.4$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\alpha_3 = 0.4$. For ICEWS-YAGO: $\alpha = 0.4$, $\alpha_1 = 0.2$, $\alpha_2 = 0.4$, $\alpha_3 = 0.4$.

3.4. Evaluation Metrics

In line with widely adopted evaluation methods in EA research, we use two metrics for evaluation:

- (1) Hits@k, measuring the percentage of correct alignments within the top k ($k = 1, 10$) matches. It is computed by $1/|\mathcal{S}| \sum_{i=1}^{|\mathcal{S}|} \mathbb{I}(\text{rank}_i < k)$, where $\mathbb{I}(\cdot)$ is the indicator function.
- (2) Mean Reciprocal Rank (MRR), reflecting the average inverse ranking of correct results. It is computed by average of the reciprocal ranks $1/|\mathcal{S}| \sum_{i=1}^{|\mathcal{S}|} \frac{1}{\text{rank}_i}$, where rank_i , $i \in \{1, \dots, |\mathcal{S}|\}$ is a set of ranking results.

Higher values in Hits@k and MRR indicate superior performance in EA.

4. Results and Discussion

4.1. Performance Comparison

Tables 2 and 3 show our results on the four datasets. It can be concluded from the results:

Table 2. Main experiment results on the four datasets. Bold: the best result; underline: the runner-up result.

Models	Hits@1	DBP15K(EN-FR) Hits@10	MRR	Hits@1	DBP-WIKI Hits@10	MRR
MTransE	0.247	0.577	0.360	0.281	0.520	0.363
BootEA	0.653	0.874	0.731	0.748	0.898	0.801
GCN-Align	0.411	0.772	0.530	0.494	0.756	0.590
RDGCN	0.873	0.950	0.901	0.974	0.994	0.980
Dual-AMN	0.954	0.994	0.970	0.983	0.996	0.991
Ours	<u>0.901</u>	<u>0.980</u>	<u>0.962</u>	<u>0.980</u>	0.993	<u>0.981</u>

Table 3. Main experiment results on the four datasets. Bold: the best result; underline: the runner-up result.

Models	Hits@1	ICEWS-WIKI Hits@10	MRR	Hits@1	ICEWS-YAGO Hits@10	MRR
MTransE	0.021	0.158	0.068	0.012	0.084	0.040
BootEA	0.072	0.275	0.139	0.020	0.120	0.056
GCN-Align	0.046	0.184	0.093	0.017	0.085	0.038
RDGCN	0.064	0.202	0.096	0.029	0.097	0.042
Dual-AMN	0.083	0.281	0.145	0.031	0.144	0.068
Ours	<u>0.081</u>	<u>0.285</u>	<u>0.148</u>	<u>0.033</u>	<u>0.146</u>	<u>0.071</u>

- In Table 2, which shows the test results of our model on two non-temporal upper datasets, the results show that our model performs consistently at a high level for most of the evaluation parameters, closely following the Dual-AMN model.
- Table 3 presents the evaluation results of our method on the time series dataset. The results clearly show that our method outperforms the other baseline models. Compared to the Dual-AMN model, our method achieves an average improvement of 0.03 in Hit@1, Hits@10 and MRR, which highlights the effectiveness and superiority of our model.
- The results clearly show that our method has a significant advantage on the temporal dataset. This advantage is due to the use of temporal features, which act as identifiers during entity alignment, enabling us to accurately determine the same entities. Analyzing the parameters α_1 , α_2 , α_3 and α , we observe that the structure-based information is

more useful in datasets with higher graph density (e.g., ICEWS-WIKI, ICEWS-YAGO). On the other hand, descriptive information of entities tends to be more advantageous in cases where the graph structure is sparser (e.g., DBP15K(EN-FR), DBP-WIKI), while alignment using semantic information proves to be more effective.

4.2. Parameter Sensitivity

We perform parameter sensitivity tests on the model, specifically analyzing the effect of the α parameter on the performance of the model. The performance of the model is evaluated using MRR. The results are shown in Figure 2. We can conclude that:

- As shown in the figure, the analysis reveals a clear trend. It is evident that each dataset has an optimal α value, beyond which the performance of the model either converges or even tends to decrease. This finding suggests that there is a critical threshold for the α parameter, beyond which further tuning may not significantly improve model performance.
- When evaluating the DBP15K(EN-FR) and DBP-WIKI datasets, we observed that the model’s MRR initially increases with increasing α values, but eventually starts to decrease. Here, α represents the weight of semantic information, indicating that increased emphasis on semantic information leads to improved performance in the early stages. However, in the later stages, the pronounced decrease in performance suggests that overemphasis on semantic information may lead to diminishing returns.
- When evaluating the ICEWS-WIKI and ICEWS-YAGO datasets, we observe a consistent trend in the performance of the model. The MRR peaks at a α value of 0.4 for these datasets, while the best overall performance for the DBP15K(EN-FR) and DBP-WIKI datasets is achieved at a α value of 0.6. This observation suggests that the ICEWS-WIKI and ICEWS-YAGO datasets have a higher dependence on the structural information of the model. This higher dependence on structural information can be attributed to the higher density and number of triples present in these datasets. Consequently, comparing structures in these datasets provides more valuable information.
- In summary, structural information is advantageous in densely structured datasets because it provides richer information about neighboring nodes and links. On the contrary, in sparsely structured datasets, semantic information is particularly helpful to compensate for the lack of structural information and to provide semantic information about characterized entities.

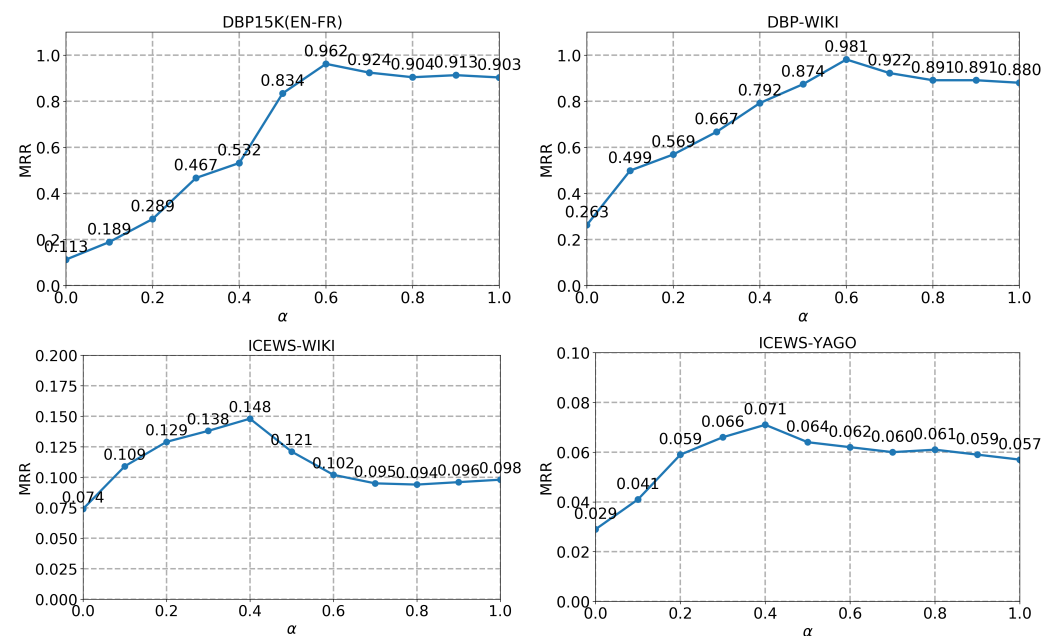


Figure 2. Results of the effect of α on model performance on four datasets.

4.3. Ablation Studies

We perform an ablation study using four datasets to evaluate the contribution of the two modules in the model. The symbol (+SSC) indicates the Semantic-based Similarity Comparison module, and the symbol (+TSC) indicates the Topology-based Similarity Comparison module. From the results, we can conclude that:

- Table 4 presents the performance of both the DBP15K(EN-FR) and DBP-WIKI datasets, demonstrating that the SSC module contributes significantly more than the TSC module to these two datasets. The SSC module exhibited a greater impact on the datasets, with an average increase of 0.318 in Hits@1, 0.244 in Hits@10 and 0.258 in MRR compared to the TSC module. This suggests that the semantic similarity module is a more effective approach for these two datasets.
- Table 5 presents the performance of the ICEWS-WIKI and ICEWS-YAGO datasets, indicating that the TSC module exhibits a significant contribution to both datasets compared to the SSC module. On average, Hits@1 improves by 0.0235, Hits@10 improves by 0.0865 and MRR improves by 0.0455. These results suggest that employing the structural similarity module is more effective for these two datasets.
- The results presented above shows that both modules contribute to the overall performance of our model. However, it is important to note that each module shows different contributions based on the KG characteristics.

Table 4. Results of ablation studies on DBP15K(EN-FR) and DBP-WIKI. Bold: the best result; underline: the runner-up result.

Models	DBP15K(EN-FR)			DBP-WIKI		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
SSC	0.793	0.810	0.892	0.849	0.930	0.890
TSC	0.560	0.678	0.692	0.445	0.574	0.573
SSC+TSC	0.901	0.980	0.962	0.980	0.993	0.981

Table 5. Results of ablation studies on ICEWS-WIKI and ICEWS-YAGO. Bold: the best result; underline: the runner-up result.

Models	ICEWS-WIKI			ICEWS-YAGO		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
SSC	0.034	0.121	0.086	0.022	0.091	0.032
TSC	<u>0.073</u>	<u>0.245</u>	<u>0.141</u>	<u>0.030</u>	<u>0.140</u>	<u>0.068</u>
SSC+TSC	0.081	0.285	0.148	0.033	0.146	0.071

5. Conclusions

This paper presents a novel entity-alignment module that combines semantic and structural comparisons. The descriptive information of an entity provides various features, semantics and characteristics, which we quantify to measure semantic similarity. Additionally, for graphs with higher structural density, we introduce a structure-based similarity-comparison module. This module assesses similarity by comparing the types of neighboring nodes and link attributes of entities across different graphs. Finally, we fuse the comparison results from both modules. Experimental results on four datasets demonstrate the effectiveness of our method. In future work, our future research aims to capture more distinguishable features for entity alignment.

Author Contributions: Conceptualization, Y.L. (Yao Liu) and Y.L. (Ye Liu); Methodology, Y.L. (Yao Liu) and Y.L. (Ye Liu); Validation, Y.L. (Yao Liu) and Y.L. (Ye Liu); Formal analysis, Y.L. (Yao Liu) and Y.L. (Ye Liu); Investigation, Y.L. (Yao Liu) and Y.L. (Ye Liu); Resources, Y.L. (Yao Liu) and Y.L. (Ye Liu); Data curation, Y.L. (Yao Liu) and Y.L. (Ye Liu); Writing – original draft, Y.L. (Yao Liu) and Y.L. (Ye Liu); Writing – review & editing, Y.L. (Yao Liu) and Y.L. (Ye Liu); Visualization, Y.L. (Yao Liu) and Y.L. (Ye Liu); Supervision, Y.L. (Yao Liu) and Y.L. (Ye Liu); Project administration, Yao Liu. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yan, Z.; Peng, R.; Wu, H. Similarity propagation based semi-supervised entity alignment. *Eng. Appl. Artif. Intell.* **2024**, *130*, 107787. [CrossRef]
2. Yang, L.; Chen, J.; Wang, Z.; Shang, F. Relation mapping based on higher-order graph convolutional network for entity alignment. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108009. [CrossRef]
3. Ranaldi, L.; Pucci, G. Knowing knowledge: Epistemological study of knowledge in transformers. *Appl. Sci.* **2023**, *13*, 677. [CrossRef]
4. Cao, J.; Fang, J.; Meng, Z.; Liang, S. Knowledge graph embedding: A survey from the perspective of representation spaces. *ACM Comput. Surv.* **2024**, *56*, 1–42. [CrossRef]
5. Ge, X.; Wang, Y.C.; Wang, B.; Kuo, C.C.J. Knowledge Graph Embedding: An Overview. *Apsipa Trans. Signal Inf. Process.* **2024**, *13*, 1–51. [CrossRef]
6. Wang, Y.; Lipka, N.; Rossi, R.A.; Siu, A.; Zhang, R.; Derr, T. Knowledge graph prompting for multi-document question answering. *AAAI Conf. Artif. Intell.* **2024**, *38*, 19206–19214. [CrossRef]
7. Yang, L.; Chen, H.; Wang, X.; Yang, J.; Wang, F.Y.; Liu, H. Two Heads Are Better Than One: Integrating Knowledge from Knowledge Graphs and Large Language Models for Entity Alignment. *arXiv* **2024**, arXiv:2401.16960.
8. Chen, C.; Zheng, F.; Cui, J.; Cao, Y.; Liu, G.; Wu, J.; Zhou, J. Survey and open problems in privacy-preserving knowledge graph: Merging, query, representation, completion, and applications. *Int. J. Mach. Learn. Cybern.* **2024**, 1–20. [CrossRef]
9. Chen, J.; Yang, L.; Wang, Z.; Gong, M. Higher-order GNN with Local Inflation for entity alignment. *Knowl.-Based Syst.* **2024**, *293*, 111634. [CrossRef]
10. Luo, S.; Yu, J. ESGNet: A multimodal network model incorporating entity semantic graphs for information extraction from Chinese resumes. *Inf. Process. Manag.* **2024**, *61*, 103524. [CrossRef]
11. Chen, Z.; Zhang, Y.; Fang, Y.; Geng, Y.; Guo, L.; Chen, X.; Li, Q.; Zhang, W.; Chen, J.; Zhu, Y.; et al. Knowledge Graphs Meet Multi-Modal Learning: A Comprehensive Survey. *arXiv* **2024**, arXiv:2402.05391.
12. Huber, M.N.; Angst, M.; Fischer, M. The Link Between Social-Ecological Network Fit and Outcomes: A Rare Empirical Assessment of a Prominent Hypothesis. *Soc. Nat. Resour.* **2024**, 1–18. [CrossRef]
13. Chen, M.; Tian, Y.; Yang, M.; Zaniolo, C. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. *arXiv* **2016**, arXiv:1611.03954.
14. Sun, Z.; Hu, W.; Zhang, Q.; Qu, Y. Bootstrapping entity alignment with knowledge graph embedding. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18, Stockholm, Sweden, 9–19 July 2018; Volume 18.
15. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26*; Curran Associates Inc.: Glasgow, UK, 2013.
16. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
17. Wang, Z.; Lv, Q.; Lan, X.; Zhang, Y. Cross-lingual Knowledge Graph Alignment via Graph Convolutional Networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 349–357.
18. Chen, Z.; Wu, Y.; Feng, Y.; Zhao, D. Integrating manifold knowledge for global entity linking with heterogeneous graphs. *Data Intell.* **2022**, *4*, 20–40. [CrossRef]
19. Mao, X.; Wang, W.; Wu, Y.; Lan, M. Boosting the speed of entity alignment 10*: Dual attention matching network with normalized hard sample mining. In *WWW '21: Proceedings of the Web Conference 2021*; Association for Computing Machinery: New York, NY, USA, 2021.
20. Xu, C.; Su, F.; Lehmann, J. Time-aware graph neural networks for entity alignment between temporal knowledge graphs. *arXiv* **2022**, arXiv:2203.02150.
21. Xu, C.; Su, F.; Xiong, V.; Lehmann, J. Time-aware Entity Alignment using Temporal Relational Attention. In Proceedings of the ACM Web Conference 2022 (WWW '22), Barcelona, Spain, Online, 26–29 June 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 788–797.
22. Cai, L.; Mao, X.; Ma, M.; Yuan, H.; Zhu, J.; Lan, M. A simple temporal information matching mechanism for entity alignment between temporal knowledge graphs. *arXiv* **2022**, arXiv:2209.09677.
23. Wang, C.; Huang, Z.; Wan, Y.; Wei, J.; Zhao, J.; Wang, P. FuAlign: Cross-lingual entity alignment via multi-view representation learning of fused knowledge graphs. *Inform. Fusion* **2023**, *89*, 41–52. [CrossRef]
24. Jiang, X.; Xu, C.; Shen, Y.; Su, F.; Wang, Y.; Sun, F.; Li, Z.; Shen, H. Rethinking gnn-based entity alignment on heterogeneous knowledge graphs: New datasets and a new method. *arXiv* **2023**, arXiv:2304.03468.
25. Tang, X.; Zhang, J.; Chen, B.; Yang, Y.; Chen, H.; Li, C. BERT-INT: A BERT-based Interaction Model For Knowledge Graph Alignment. *Interactions* **2020**, *100*, e1.
26. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.* **2024**, *15*, 1–45. [CrossRef]

27. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*; Curran Associates Inc.: Glasgow, UK, 2019.
28. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. Preprint. 2018. Available online: <https://paperswithcode.com/paper/improving-language-understanding-by> (accessed on 20 May 2024).
29. Dai, Z.; Yang, Z.; Yang, Y.; Carbonell, J.; Le, Q.V.; Salakhutdinov, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv* **2019**, arXiv:1901.02860.
30. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
31. Jiang, X.; Shen, Y.; Shi, Z.; Xu, C.; Li, W.; Li, Z.; Guo, J.; Shen, H.; Wang, Y. Unlocking the Power of Large Language Models for Entity Alignment. *arXiv* **2024**, arXiv:2402.15048.
32. Lynch, C.J.; Jensen, E.; Munro, M.H.; Zamponi, V.; Martinez, J.; O'Brien, K.; Feldhaus, B.; Smith, K.; Reinhold, A.M.; Gore, R. GPT-4 Generated Narratives of Life Events using a Structured Narrative Prompt: A Validation Study. *arXiv* **2024**, arXiv:2402.05435.
33. Sun, Z.; Zhang, Q.; Hu, W.; Wang, C.; Chen, M.; Akrami, F.; Li, C. A benchmarking study of embedding-based entity alignment for knowledge graphs. *arXiv* **2020**, arXiv:2003.07743.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.