electronics

MDPI

Article

A Multi-Information Fusion Unsupervised Entity Alignment Model for Knowledge Graphs in Oil and Gas Pipeline Safety

Wangweiyi Shan [1], Heng Duan [1,*], Weichun Chang [2], Kewen Li [1] and Guangyue Zhou [1]

[1] Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China
[2] PipeChina Science and Technology Research Institute, Binhai, Tianjin 610500, China
* Correspondence: z23070067@s.upc.edu.cn

Abstract

Targeting the joint challenges posed by sparse graph topology, limited semantic expressiveness, and scarce annotation resources that commonly afflict knowledge graphs in the oil and gas pipeline safety domain, this paper presents a Multi-Information Fusion Unsupervised Entity Alignment model (MIF-UEA). The proposed method constructs high-quality initial alignment pairs by integrating multi-source similarity computation with a structure-aware seed generation mechanism and performs representation learning by fusing structural features and semantic attribute information. Furthermore, a pseudo-label augmentation and denoising strategy is introduced to enhance the effectiveness of self-training. Finally, entity matching is achieved through an optimal transport model. Experimental results confirm that MIF-UEA surpasses existing baselines across both the specialized oil and gas pipeline safety dataset and multiple general-domain benchmarks, demonstrating its effectiveness and generalization capability.

Keywords: entity alignment; knowledge graph; unsupervised learning; multi-information fusion; oil and gas pipeline safety

1. Introduction

Knowledge graphs (KGs) provide a structured paradigm for organizing entities, along with their interrelations, and have been extensively adopted across a broad spectrum of domains, including but not limited to recommendation systems, question answering, and industrial safety management. Constructing knowledge graphs for oil and gas pipeline safety can comprehensively integrate oil and gas pipeline safety information, identify potential risk points, and provide strong support for the safe operation of oil and gas pipelines [1,2]. However, due to the diversity of data sources and limited coverage, a single knowledge graph often fails to fully capture domain knowledge in complex scenarios. Therefore, fusing multi-source knowledge graphs has become an important means of enhancing knowledge completeness. Entity alignment (EA) [3–5], as a core technology for multi-source knowledge graph fusion, aims to identify entities in different knowledge graphs that refer to the same real-world object, thereby achieving cross-graph semantic fusion and mutual knowledge enrichment.

As representation learning has advanced in recent years, embedding-based approaches have gradually emerged as the prevailing paradigm for entity alignment. These methods can be broadly divided into two categories: those built upon TransE [6] and those

Academic Editor: Ioannis Hatzilygeroudis

Received: 25 March 2026
Revised: 23 April 2026
Accepted: 28 April 2026
Published: 7 July 2026

leveraging Graph Neural Networks (GNNs) [7,8]. TransE-based approaches project relational triples of a knowledge graph into a low-dimensional continuous space, where the translation principle enforces that the vector sum of a head entity and its associated relation closely approximate the corresponding tail-entity vector. Such methods are valued for their straightforward formulation and favorable computational scalability. However, these methods primarily rely on local triple constraints and have difficulty in fully capturing the global structural information of entities. To address this limitation, GNN-based methods perform multi-layer information propagation and neighborhood aggregation on graph structures, effectively encoding the topological structural features of entities and significantly improving alignment performance through multi-hop neighborhoods and attention mechanisms. Furthermore, researchers have recently begun to introduce attribute information and textual semantic information, enhancing entity representation capabilities by fusing entity attributes or contextual descriptions, thereby improving the accuracy of entity alignment.

Although the above methods have achieved satisfactory results on public benchmark datasets, they still face numerous challenges in practical applications within the oil and gas pipeline safety domain. First, knowledge graphs in this domain typically exhibit significant structural sparsity, with a large number of entities connected to only a few neighbor nodes, making it difficult for structure propagation-based representation learning methods to fully leverage their potential. This sparsity originates from the Bow-Tie analysis model that underlies domain knowledge graph construction, which organizes safety knowledge along the causal chain of "hazard sources $\rightarrow$ preventive barriers $\rightarrow$ top event $\rightarrow$ mitigating barriers $\rightarrow$ consequences". The Bow-Tie model involves only a handful of predefined relation types (e.g., "causes", "prevents", and "mitigates"), resulting in each entity typically connecting to only two to three neighbors. In the oil and gas pipeline safety entity alignment dataset constructed in this study, the average number of relational triples per entity is approximately 1.1, whereas in general-domain datasets, this value is approximately 3. To illustrate, in a typical domain knowledge graph, the entity "external corrosion perforation" (a risk event) participates in only three relational triples: $\langle$anti-corrosion coating inspection, prevention, anti-corrosion coating aging$\rangle$, $\langle$anti-corrosion coating aging, triggering, external corrosion perforation$\rangle$, and $\langle$external corrosion perforation, consequence, fire and explosion$\rangle$, whereas entities in general-domain knowledge graphs (e.g., "Beijing" in DBpedia) can be connected to hundreds of entities through dozens of relation types. This difference stems from the fact that general-domain knowledge graphs extract entities from multiple open data sources with naturally multi-dimensional associations, while oil and gas pipeline safety knowledge graphs derive entities from domain-specific Bow-Tie analysis materials with relation types restricted to a small set of predefined semantic roles (only four relation types in our dataset). Such low connectivity density makes it difficult for structure propagation-based methods to learn discriminative entity embeddings. It is worth noting that structural sparsity is not unique to this domain; other industrial safety fields that construct knowledge graphs based on the Bow-Tie model or similar tree-structured risk analysis frameworks also face comparable challenges. Second, entities share highly similar attribute types and primarily differ in their attribute values, making it difficult for alignment methods that rely on attribute types or structural patterns to effectively distinguish entities. Moreover, in the overall knowledge graph formed by concatenating multiple sub-domain knowledge graphs, the entity similarity distribution often exhibits imbalance, further increasing alignment difficulty. Finally, due to the high degree of specialization in the oil and gas pipeline safety domain, high-quality alignment annotations typically depend on expert knowledge, which is costly to obtain and difficult to scale, limiting the practical application of supervised methods.

To address the above issues, this paper proposes a Multi-Information Fusion Unsupervised Entity Alignment model (MIF-UEA) for oil and gas pipeline safety knowledge graphs. The proposed method first fuses entity name semantics, attribute textual information, and string similarity to generate high-quality initial alignment seeds. In the representation learning phase, structural information and attribute information are jointly modeled to enhance entity representation capabilities. During training, a pseudo-label augmentation and denoising strategy is introduced to improve model robustness. Finally, an optimal transport formulation is leveraged to obtain globally coherent entity correspondences, which further elevates the holistic alignment quality.

2. Related Work

2.1. Structure-Based Entity Alignment Methods

Entity alignment methods that rely on graph structure exploit the topological patterns embedded in knowledge graphs as their primary source of supervisory signal. Within this category, TransE-based alignment techniques first acquire low-dimensional embeddings for entities and relations in each individual knowledge graph independently and subsequently project heterogeneous embedding spaces into a shared vector space via cross-graph alignment transformations, thereby enabling entity-level matching across different graphs. MTransE [9] constructs independent embedding spaces for different knowledge graphs and learns cross-space transformation matrices based on already-aligned entities to achieve the mapping and alignment of entity representations. BootEA [10] further introduces a bootstrapping mechanism, continuously expanding high-confidence alignment seeds during the training process to improve model performance, and mitigates error propagation through an alignment editing strategy.

Although TransE-based methods offer advantages such as simple modeling and high computational efficiency, they primarily rely on local triple constraints and neglect the global structural information of entities within the graph. To better integrate neighbor information, researchers have introduced graph neural networks to construct alignment models. AliNet [11] effectively alleviates entity heterogeneity between different knowledge graphs through a gated multi-hop neighborhood aggregation mechanism and improves embedding quality in combination with relation modeling. MRAEA [12] starts from the meta-semantics of in-edge neighbors, out-edge neighbors, and their relations and combines a bidirectional iterative strategy to gradually expand alignment seeds during the training process, thereby enhancing the discriminative ability of entity representations. Dual-AMN [13] addresses the issues of high model complexity and low negative sampling efficiency by modeling aligned entities as a special relation, replacing partial cross-graph node interactions with proxy vectors, and introducing a normalized hard sample mining loss to improve training efficiency and alignment performance. RHGN [14] distinguishes the semantic spaces of relations and entities through relation-gated convolution and uses soft relation alignment labels to identify mutually similar relations between different knowledge graphs.

Structure-based methods can effectively leverage the topological relationships in knowledge graphs and typically achieve good performance on datasets with relatively complete structures and rich neighborhood information. However, these methods have a strong dependence on graph structure, and their performance is often limited in scenarios with sparse structures or insufficient relational information. Beyond knowledge graph alignment, graph neural network architectures have also demonstrated effectiveness in related industrial safety tasks, such as equipment fault diagnosis [15], where graph structures are constructed to capture spatial relationships among discriminant features, further highlighting the potential of graph-based methods in industrial domains.

2.2. Semantic Information-Enhanced Entity Alignment Methods

To compensate for the limitations of structural information, researchers have recently begun to introduce semantic information to enhance entity representation capabilities. Such methods typically fuse entity names, attribute information, or textual descriptions to characterize semantic entity featuresfrom multiple perspectives. MultiKE [16] is one of the representative methods that first introduced the multi-view concept into entity alignment. This model decomposes entity representation into three complementary perspectives—namely, a name view, a relation view, and an attribute view—and performs entity alignment via collaborative learning across all views. This method demonstrates that structural information alone is insufficient to fully characterize entity semantics, and fusion from multiple perspectives helps improve alignment accuracy. AttrGNN [17] further strengthens the modeling capability of attribute information by simultaneously processing attribute triples and relation triples within a unified framework and dynamically learns the importance of different attributes and attribute values through graph partitioning and attribute value encoders, thereby enhancing the role of attribute semantics in the alignment process. RREA [18] effectively enhances the expressive power of relational information through relational reflection transformations and dual-view embedding modeling and improves alignment performance in combination with entity name information. CAEA [19] approaches from the perspective of concept semantics, constructing relative and independent concept representations by aggregating entity relation and attribute information, and introduces a concept-aware graph convolutional network to strengthen key entities and their conceptual semantic expressions while combining BERT for semantic enhancement of attribute text to adapt to application scenarios with complex semantics and distinct conceptual hierarchies. In addition to the above methods, some studies have begun to focus on reducing the dependence on manual annotations. SelfKG [20] initializes entity semantics using LaBSE and optimizes the distribution of entity representation through relative similarity metrics and contrastive learning strategies, achieving entity alignment without relying on a large number of alignment labels. UDCEA [21] combines machine translation with multilingual pre-trained encoders to construct multi-view entity representations and fuses global and local information to optimize bipartite graph matching, thereby improving entity alignment performance in cross-lingual scenarios.

Methods that incorporate semantic information can partially compensate for the limitations of structural information; however, their discriminative ability remains limited in the oil and gas pipeline safety domain. Therefore, how to achieve more effective modeling of both structural information and multi-source semantic information remains a key issue to be addressed for entity alignment in this domain.

2.3. Knowledge Graphs in the Oil and Gas Pipeline Safety Domain

In recent years, knowledge graph technologies have been increasingly applied to the oil and gas pipeline safety domain. Wu [1] constructed a knowledge graph for geological disaster risk management of oil and gas pipelines, integrating multi-source geological hazard data to support risk assessment and decision-making. Chen et al. [2] proposed a knowledge graph-based early warning method for accident evolution at overseas natural gas pipeline stations under harsh environmental conditions, using Bi-LSTM-CRF to extract causal relationships from accident reports. Bai et al. [22] integrated knowledge graphs with DEMATEL and Bayesian networks to establish a data-driven risk assessment model for natural gas pipelines, replacing traditional expert-dependent approaches with automated causal network extraction from accident reports. Chen et al. [23] constructed the first knowledge graph for long-distance oil and gas pipeline emergency cases and employed graph convolutional networks to improve emergency task recommendation accuracy.

Simone et al. [24] proposed a methodology to extract safety knowledge from industrial near-miss reports and construct knowledge graphs, demonstrating its applicability to oil-refinery plant data. These studies demonstrate the value of knowledge graphs in organizing domain safety knowledge. However, existing research has primarily focused on knowledge graph construction and downstream applications such as risk warning, accident analysis, and emergency recommendation, while entity alignment for multi-source knowledge graph fusion remains largely unexplored. Furthermore, no publicly available entity alignment benchmark dataset exists for this domain. This research gap motivates both the construction of the Pipe-DPMEA dataset and the development of the MIF-UEA model proposed in this paper.

3. Problem Definition

A knowledge graph is formally represented as a six-tuple, i.e., $G = \langle E, R, A, V, X, Y \rangle$, in which E, R, A, and V correspond to the sets of entities, relations, attributes, and attribute values, respectively. The subset expressed as $X \subseteq E \times R \times E$ captures the relational structure among entities; each element ($x \in X$) takes the form of a triple $\langle h, r, t \rangle$, where $h \in E$ is the head entity, $r \in R$ is the relation, and $t \in E$ is the tail entity. Analogously, $Y \subseteq E \times A \times V$ encodes the attribute-level information of entities; each element ($y \in Y$) is a triple $\langle h, a, v \rangle$, where $h \in E$ refers to the entity, $a \in A$ to the attribute, and $v \in V$ to the associated attribute value. Given two knowledge graphs ($G_1 = \langle E_1, R_1, A_1, V_1, X_1, Y_1 \rangle$ and $G_2 = \langle E_2, R_2, A_2, V_2, X_2, Y_2 \rangle$), entity alignment seeks to identify the set of cross-graph entity correspondences ($S = \{(e_i, e_j) \in E_1 \times E_2 \mid e_i \equiv e_j\}$, where $\equiv$ indicates that e_i and e_j refer to the same real-world entity). An illustrative example is given in Figure 1.

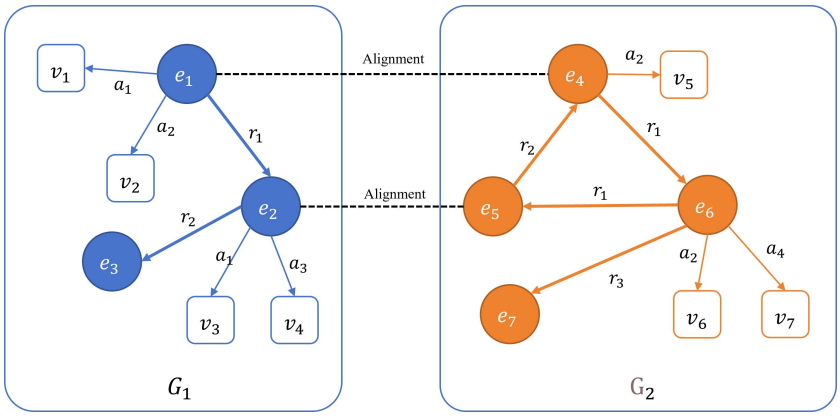

Figure 1. An example of entity alignment.

4. Method

This section provides a detailed account of each component in the proposed method. Figure 2 depicts the end-to-end model architecture. The model is organized around four principal modules: (1) the multi-information fusion unsupervised seed generation module, (2) the structure–attribute fusion representation learning module, (3) the pseudo-label augmentation and denoising module, and (4) the optimal transport-based entity-matching module.

First, a comprehensive similarity matrix is constructed based on name semantics, attribute semantics, and string similarity, combined with a structure-aware mechanism to screen high-confidence initial seeds. Subsequently, low-dimensional embedding representations of entities are learned through the structure–attribute encoder. Throughout the iterative self-training procedure, pseudo-labels selected with high confidence serve to progressively augment the training corpus while filtering out noisy samples, which,

in turn, drives sustained refinement of the learned embedding space. In the final stage, entity alignment is cast as an optimal transport problem, and a global matching solution is derived to produce the definitive alignment outcomes.

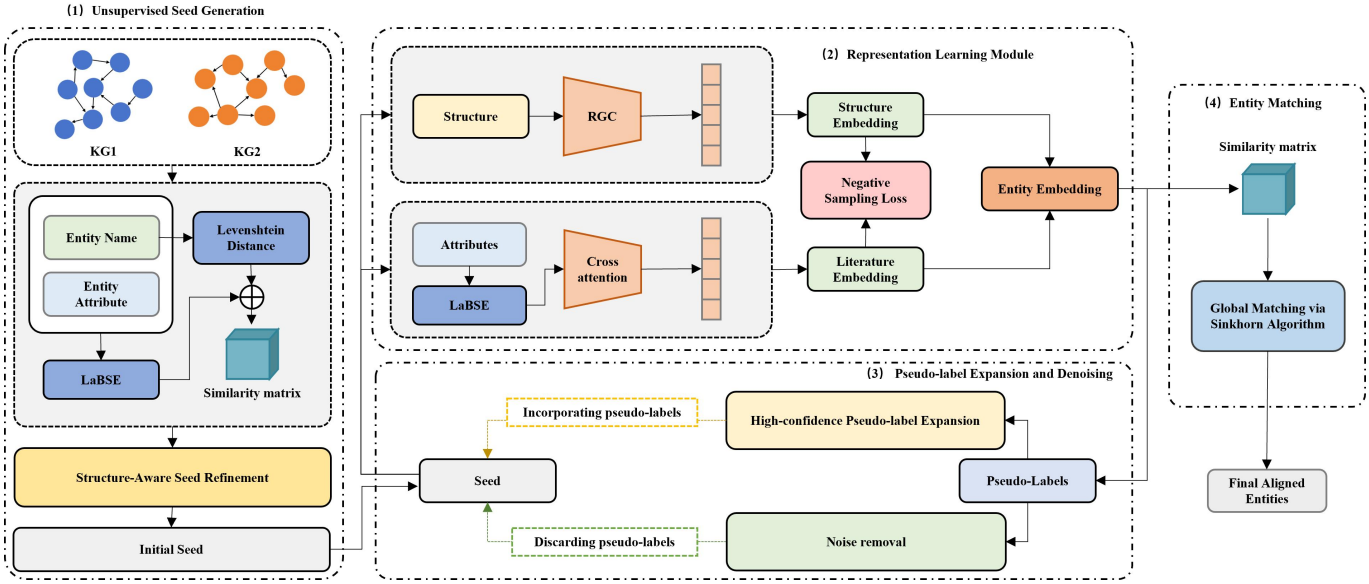

Figure 2. Overall framework of the MIF-UEA model. The model consists of four modules: (1) unsupervised seed generation, (2) structure–attribute fusion representation learning, (3) pseudo-label augmentation and denoising, and (4) entity matching.

4.1. Multi-Information Fusion Unsupervised Seed Generation Module

In unsupervised entity alignment tasks, the quality of initial alignment seeds directly determines the convergence behavior and final accuracy of subsequent self-training iterations. To this end, the MIF-UEA model designs a seed generation mechanism that fuses multi-source semantics with structure-aware penalties, extracting high-quality, evenly distributed initial supervision signals under annotation-free conditions.

4.1.1. Multi-Source Semantic and Character Feature Modeling

To fully characterize the semantic and character-level similarity between entities, similarity matrices are constructed from three perspectives: name semantics, attribute semantics, and string edit distance. Specifically, the model first uses the pre-trained LaBSE language model [25] to map entity names into a unified continuous vector space and computes their cosine similarity to generate the semantic name matrix S_{name}. Furthermore, the multi-dimensional attributes of entities are modeled by concatenating and average pooling the LaBSE-encoded attribute name and attribute value vectors, then computing pairwise similarities to construct the semantic attribute matrix S_{attr}. Additionally, to compensate for the insensitivity of the continuous semantic space to character-level local variations (such as word-order adjustments and abbreviations), the Levenshtein edit distance [26] is introduced to construct the string similarity matrix S_{edit}. A weighted linear fusion strategy is adopted to construct the final semantic similarity matrix. The comprehensive similarity matrix is defined as follows:

$$S_{\text{final}} = \alpha_1 S_{\text{name}} + \alpha_2 S_{\text{attr}} + \alpha_3 S_{\text{edit}} \tag{1}$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$, $\alpha_1, \alpha_2, \alpha_3 \geq 0$, and α_1, α_2, α_3 are hyperparameters that control the contribution proportions of the semantic name matrix, semantic attribute matrix, and edit distance matrix, respectively.

4.1.2. BFS-Based Structure-Aware Penalty Mechanism

The oil and gas pipeline safety domain knowledge graphs to be aligned are typically formed by fusing knowledge graphs from multiple sub-domains, resulting in highly non-uniform similarity distributions within the graph. If simple greedy matching is performed based solely on S_{final}, the initial seeds tend to cluster excessively in locally high-density entity clusters, leading to a lack of supervision signals in other sparse regions of the graph, which, in turn, biases the model's learning of global topological features.

To mitigate this problem, we incorporate Cross-domain Similarity Local Scaling (CSLS) [27], and a BFS-based structure-aware penalty mechanism is further designed, as detailed in Algorithm 1.

Algorithm 1 Structure-Aware Multi-Source Similarity Seed Generation Algorithm

Require: Fused entity similarity matrix S_{final}, KG1 entity set V_1, KG2 entity set V_2, KG1 and KG2 adjacency lists Adj_1, Adj_2, seed ratio β, semantic weight coefficient δ, BFS maximum depth $d_{\max}$
Ensure: Initial alignment seed set $\mathcal{S}$
$N \leftarrow \lfloor |V_1| \times \beta \rfloor$
$S' \leftarrow \text{CSLS}(S_{\text{final}})$
Initialize $\mathcal{S} \leftarrow \emptyset$
Initialize matched sets $U_1 \leftarrow \emptyset$, $U_2 \leftarrow \emptyset$
Initialize structural influence arrays $\text{inf}_1[i] \leftarrow 0$, $\text{inf}_2[j] \leftarrow 0$
For each entity $i \in V_1$, select Top-K candidate indices from S' to build candidate pool $\text{Cand}[i]$
while $|\mathcal{S}| < N$ **do**
  best_score $\leftarrow -\infty$, best_pair $\leftarrow$ None
  for each $i \in V_1$ and $i \notin U_1$ **do**
    for each $j \in \text{Cand}[i]$ and $j \notin U_2$ **do**
      Compute cumulative structural penalty: penalty $\leftarrow \text{inf}_1[i] + \text{inf}_2[j]$
      Compute composite score: score $\leftarrow \delta \cdot S'[i,j] + (1-\delta)/(1+\text{penalty})$
      if score $>$ best_score **then**
        best_score $\leftarrow$ score, best_pair $\leftarrow (i,j)$
      end if
    end for
  end for
  if best_pair is None **then**
    break
  end if
  Add best_pair to $\mathcal{S}$; $U_1 \leftarrow U_1 \cup \{i\}$; $U_2 \leftarrow U_2 \cup \{j\}$
  Call Incremental_BFS$(i, \text{Adj}_1, \text{inf}_1, d_{\max})$: for neighbor v at depth $d \le d_{\max}$, accumulate $\text{inf}_1[v] \leftarrow \text{inf}_1[v] + 1/d$
  Call Incremental_BFS$(j, \text{Adj}_2, \text{inf}_2, d_{\max})$: for neighbor v at depth $d \le d_{\max}$, accumulate $\text{inf}_2[v] \leftarrow \text{inf}_2[v] + 1/d$
end while
return $\mathcal{S}$

This algorithm achieves synergistic optimization of semantic information and structural information. Compared with seed generation methods based solely on similarity ranking, this mechanism effectively mitigates the excessive clustering of seeds in the graph structure, enabling the initial supervision signals to exhibit a more balanced distribution across the graph space while maintaining a reasonable quality of the initial alignment seed set.

The computational cost of Algorithm 1 is primarily determined by two components: (1) the candidate selection loop, whose worst-case complexity is $O(N \times |V_1| \times K)$, where $N = \lfloor |V_1| \times \beta \rfloor$ denotes the number of seeds to be selected, and (2) the incremental BFS updates, with a total cost of $O(N \times \bar{d}^{\,d_{\max}})$, where $\bar{d}$ is the average node degree. In practice,

since the BFS depth ($d_{\max}$) is typically set to a small value and the graph structure is generally sparse, the computational overhead of this algorithm remains manageable, even for larger knowledge graphs.

Furthermore, to ensure that entity pairs with extremely high semantic consistency are not delayed or omitted due to the structural dispersion constraint, entity pairs with extremely high semantic consistency are directly added to the initial alignment seed set after the structure-aware screening mechanism is completed—specifically, in the corrected comprehensive similarity matrix ($S'[i,j]$), for entity pairs satisfying the bidirectional nearest neighbor constraint, i.e., simultaneously satisfying

$$j = \arg\max_{j \in E_2} S'[i,j] \tag{2}$$

and

$$i = \arg\max_{u \in E_1} S'[u,j]. \tag{3}$$

Whenever the confidence score ($\mathrm{Conf}(i,j) = S'[i,j]$) surpasses the predefined similarity threshold (τ), the corresponding entity pair is immediately incorporated into the initial seed set. Formally, the resulting seed set (Seed) is defined as follows:

$$\mathrm{Seed} = \mathcal{S} \cup \{(i,j) \mid \mathrm{Conf}(i,j) \geq \tau\} \tag{4}$$

4.2. Structure–Attribute Fusion Representation Learning

To fully characterize entity features in knowledge graphs, we design a structure–attribute encoder to learn entity representations from both structural information and attribute information perspectives.

4.2.1. Structural Information Embedding

In knowledge graphs, the meaning of a relation is closely related to the two entities connected at its head and tail. Inspired by prior work [28,29], to address the structural sparsity of knowledge graphs in the oil and gas pipeline safety domain, we dynamically construct relation representations from entity representations. Additionally, following previous work [11,30], inverse relations are also added to the knowledge graph. The structural representation of a relation is defined as the mean of the embeddings of its corresponding head-entity set and tail-entity set:

$$\mathbf{r} = \frac{1}{2}\left(\frac{1}{|H_r|}\sum_{h \in H_r}\mathbf{e}_h + \frac{1}{|T_r|}\sum_{t \in T_r}\mathbf{e}_t\right) \tag{5}$$

where $\mathbf{e}_h$ and $\mathbf{e}_t$ denote the learned vector representations associated with the head entity (h) and the tail entity (t), respectively. This vector encodes the averaged semantics of the associated head and tail entities, thereby reflecting the characteristics of the relation within the current graph structure.

To avoid drastic fluctuations in relation representations during training, the relation embedding is computed using Equation (5) in the first training round and subsequently updated using an exponential moving-average strategy:

$$\mathbf{r}_{\mathrm{new}} = (1-\varepsilon)\mathbf{r}_{\mathrm{old}} + \varepsilon\mathbf{r} \tag{6}$$

where $\varepsilon \in (0,1)$ is the relation update ratio hyperparameter that controls the influence of new structural information on the relation embedding. A larger ε tends to quickly follow entity-embedding changes, while a smaller ε maintains embedding stability. Since a single relation can connect to a wide range of head and tail entities within the graph,

exhaustively recomputing the embedding of each relation at every training step would be computationally intractable. Therefore, a random sampling mechanism is adopted, where a fixed number of head–tail entity pairs are randomly selected for each relation to perform updates.

In the aggregation phase, Relation-Gated Convolution (RGC) is employed to aggregate neighbor entities and their relational information. A key merit of RGC is that it maintains distinct semantic subspaces for relations and entities, which helps suppress noise propagation during aggregation. Concretely, we use $\mathbf{h}_i^{(k)}$ to denote the embedding of entity e_i produced by the k-th layer. The output of the subsequent $(k+1)$-th layer, i.e., $\mathbf{h}_i^{(k+1)}$, is computed as follows:

$$\mathbf{g}_{ij}^{(k)} = \text{Sigmoid}\left(\text{ReLU}\left(\mathbf{W}_r^{(k)}\mathbf{r}_{ij}^{(k)} + \mathbf{b}_r^{(k)}\right)\right) \tag{7}$$

$$\mathbf{h}_i^{(k+1)} = \tanh\left(\sum_{j\in\mathcal{N}_i} \mathbf{W}_e^{(k)}\mathbf{h}_j^{(k)} \odot \mathbf{g}_{ij}^{(k)}\right) \tag{8}$$

Here, $\mathcal{N}_i$ refers to the neighborhood of entity i. The $\mathbf{r}_{ij}^{(k)}$ vector encodes the relation along the edge from entity j to entity i, and the three nonlinear functions (tanh, Sigmoid, and ReLU) serve as activation operators at different stages of the computation. The $\mathbf{W}_r^{(k)}$ and $\mathbf{b}_r^{(k)}$ parameters are the trainable projection matrix and bias associated with the relation gate at layer k, whose output is $\mathbf{g}_{ij}^{(k)}$. The $\odot$ symbol denotes element-wise multiplication, and $\mathbf{W}_e^{(k)}$ is the trainable entity projection matrix at layer k. Upon completion of training, the final structural embedding ($\mathbf{h}_i^{\text{str}}$) for each entity ($e_i$) is derived. Additionally, following [14], a cross-graph structure is constructed, and entity embeddings from the original graph and the cross-graph are exchanged at each intermediate layer to enable more efficient information propagation between knowledge graphs.

4.2.2. Semantic Attribute Attention Aggregation

Beyond graph structure, attribute information constitutes another critical dimension for capturing the semantic characteristics of entities. In the oil and gas pipeline safety domain, different entities have relatively similar attribute types, while attribute values contain more discriminative semantic information. Therefore, we employ the pre-trained LaBSE language model [25] to jointly encode attribute names and attribute values and use an attention mechanism to perform weighted aggregation of attribute information. Specifically, for the j-th piece of attribute information (a_{ij}, v_{ij}) of entity e_i, LaBSE encoding is applied as follows:

$$\mathbf{q}_{ij} = \text{LaBSE}(a_{ij}), \quad \mathbf{k}_{ij} = \text{LaBSE}(v_{ij}) \tag{9}$$

Considering that different attributes contribute differently to entity semantics, an attribute-level attention mechanism is used to compute weights. The attribute embedding ($\mathbf{h}_i^{\text{attr}}$) of entity e_i is computed as follows:

$$\beta_{ij} = \frac{\exp\left(\mathbf{W}_Q\mathbf{q}_{ij} \cdot \mathbf{W}_K\mathbf{k}_{ij}/\sqrt{d}\right)}{\sum_{m=1}^{n}\exp\left(\mathbf{W}_Q\mathbf{q}_{im} \cdot \mathbf{W}_K\mathbf{k}_{im}/\sqrt{d}\right)} \tag{10}$$

$$\mathbf{h}_i^{\text{attr}} = \tanh\left(\mathbf{W}_o\sum_{j=1}^{n}\beta_{ij}\mathbf{W}_V\mathbf{k}_{ij}\right) \tag{11}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d \times d}$ and $\mathbf{W}_o \in \mathbb{R}^{d \times d_h}$ are learnable parameter matrices, d_h is the attribute-embedding dimension, β_{ij} represents the weight of the j-th attribute value of entity e_i in its overall attribute representation, and tanh is the activation function.

4.2.3. Fusion of Structural and Semantic Representations

Once both the topology-derived embedding and the attribute-level semantic embedding have been computed, they are fused via straightforward vector concatenation to yield the final entity representation ($\mathbf{h}_i$):

$$\mathbf{h}_i = \mathbf{h}_i^{\text{str}} \| \mathbf{h}_i^{\text{attr}} \tag{12}$$

Based on this fused representation, entity similarities can be further computed for subsequent seed generation and entity alignment tasks.

4.3. Pseudo-Label Augmentation and Denoising

In unsupervised entity alignment frameworks, the initial seed set is typically limited in scale, making it difficult to support the model in learning stable cross-graph mapping relationships. Therefore, the training samples need to be continuously expanded through a self-training mechanism [31]. However, the pseudo-label augmentation process is prone to introducing erroneous matches, and error accumulation can occur during iterations, thereby affecting model performance. To this end, we design a pseudo-label augmentation and denoising module, including a high-confidence pseudo-label augmentation strategy and a noise removal strategy, to improve the scale of training samples while ensuring data quality.

4.3.1. High-Confidence Pseudo-Label Augmentation Strategy

First, a coarse-grained filtering strategy is adopted. Given an unmatched entity from the source graph, the retrieval scope is restricted to those target graph candidates that simultaneously rank within the Top-k ($k = 20$, Top-20) by similarity score and satisfy the minimum threshold (τ), effectively narrowing the candidate space for downstream assessment. After obtaining the preliminary candidate set, directly selecting the highest-scoring entity as the pseudo-label still carries a significant risk of mismatching. When a source entity is projected into a dense cluster in the target graph, it may produce extremely high similarity scores with multiple target entities. Suppose the score of the Top-1 entity is s_i^1 and the score of the Top-2 entity is s_i^2. If the two are very close (i.e., $s_i^1 \approx s_i^2$), this means that the embedding vectors of these two entities are very close in space. Forcibly introducing such a Top-1 entity pair lacking discriminative power as a pseudo-label can easily produce erroneous labels. To address this issue, the alignment certainty of an entity is quantitatively computed by comprehensively calculating the following metric:

$$c_i = \frac{1}{2}(s_i^1 - s_i^2) + \frac{1}{2}\left(\frac{k\sum_{j=1}^{k} p_{ij}^2 - 1}{k-1}\right) \tag{13}$$

where $s_i^1 - s_i^2$ denotes the margin separating the highest- and second-highest-ranked candidates in terms of similarity, which serves as an indicator of the relative discriminative confidence. The second term measures the sharpness of the global distribution; the sharper the distribution (i.e., the probability mass is extremely concentrated on the Top-1 entity), the closer this squared sum approaches 1 and the higher the overall score of this term. Conversely, if the model exhibits high uncertainty in a dense cluster, causing the probability distribution to become uniform, the score of this term approaches 0. The two terms in c_i share the same $[0, 1]$ value range. The margin term satisfies $s_i^1 - s_i^2 \in [0, 1]$, as the

similarity scores are cosine similarities normalized to $[0, 1]$. For the sharpness term, since $\sum_{j=1}^{k} p_{ij}^2 \in [1/k,\ 1]$, it follows that $k \cdot \sum p_{ij}^2 \in [1,\ k]$. After applying the normalization factor $(\frac{1}{k-1}(k \sum p_{ij}^2 - 1))$, this term also falls within $[0, 1]$. We therefore directly adopt equal-weight averaging for their combination. When $k = 1$, $c_i = s_i^1$. p_{ij} is the relative probability of candidate entity j computed through the Softmax function with a temperature hyperparameter of T:

$$p_{ij} = \frac{\exp(s_{ij}/T)}{\sum_{z=1}^{k} \exp(s_{iz}/T)} \tag{14}$$

The primary purpose of the alignment certainty score (c_i) is to filter ambiguous matches rather than dominate the final similarity ranking. Therefore, its weight should be significantly lower than the original similarity score. The final similarity is computed as follows:

$$s_{\text{final}} = \lambda_c \cdot c_i + (1 - \lambda_c) \cdot s_i \tag{15}$$

4.3.2. Noise Removal Strategy

In the early stages of iteration, since the entity-embedding representations are not yet fully aligned, the model is highly prone to producing erroneous pseudo-alignment seeds. If erroneous alignment samples are retained in the training set for extended periods, they will continuously interfere with the embedding space and may even cause the model to deviate from the true alignment distribution. Therefore, employing a noise removal strategy is of significant importance. However, if noise evaluation is performed only at the t-th round by comparing the similarity score computed in the current round with the similarity threshold (τ) for removal, the evaluation is highly susceptible to misjudgment caused by single-iteration fluctuations. To address this issue, instead of directly comparing the current round's $\text{Score}_{\text{joint}}^{(t)}$ with the similarity threshold (τ), the model updates the time-smoothed composite score ($\text{SEMA}^{(t)}$) for the pseudo-label. The corresponding update rule is formulated as follows:

$$\text{SEMA}^{(t)} = \mu \cdot \text{SEMA}^{(t-1)} + (1 - \mu) \cdot \text{Score}_{\text{joint}}^{(t)} \tag{16}$$

where $\mu \in (0, 1)$ is the time-smoothing coefficient that controls the trade-off between the historical score and the current score. If the smoothed score of an aligned entity pair falls below the similarity threshold (τ), i.e., $\text{SEMA}^{(t)} < \tau$, the pseudo-label is removed from the training set.

4.4. Optimal Transport-Based Entity Matching

After obtaining the entity-embedding representations of the oil and gas pipeline safety knowledge graph, the embedding space needs to be converted into concrete entity alignment results through similarity computation. Since the knowledge graphs in practical business scenarios often exhibit scale imbalance and contain a large number of dangling entities that have no corresponding entity in the target graph, a matching strategy combining fixed threshold filtering with augmented Sinkhorn optimal transport is adopted.

To begin with, cosine similarity scores are computed between every pair of test entity embeddings drawn from the two graphs to be aligned, producing the raw similarity matrix ($\mathbf{S}$). If the maximum similarity score of a given row or column of the similarity matrix is still below the truncation threshold (τ'), the corresponding node is identified as a dangling entity and directly removed from the current matching pool.

After the initial threshold truncation, the numbers of remaining source and target entities to be matched are M and N, respectively. Since, typically, $M \neq N$, the truncated similarity matrix ($\mathbf{S}' \in \mathbb{R}^{M \times N}$) is non-square. Let $Q = \max(M, N)$, and the original

similarity matrix ($\mathbf{S}'$) is padded and extended to a $Q \times Q$ square matrix ($\hat{\mathbf{S}}$). The element values of newly padded virtual rows or virtual columns are uniformly set to the minimum similarity value in the current matrix ($\mathbf{S}'$). Entities matched to these padded nodes are treated as dangling entities. With this formulation in place, the entity alignment task naturally reduces to an optimal transport problem [32–34], which is then solved over the square cost matrix via the sparse Sinkhorn algorithm [20,35] to derive the definitive alignment results.

5. Experiments

5.1. Datasets

To evaluate model performance, experiments are conducted on the following datasets: a self-constructed dataset for the oil and gas pipeline safety domain (Pipe-DPMEA) and publicly available general-domain datasets [19] (SRPRS_EN-FR-15K, SRPRS_EN-DE-15K, and SRPRS_D-Y-15K).

SRPRS_EN-FR-15K, SRPRS_EN-DE-15K, and SRPRS_D-Y-15K are publicly available general-domain datasets extracted from DBpedia and YAGO3, featuring well-formed topological structures and fully aligned entities, which enable performance evaluation under ideal conditions. Among them, SRPRS_EN-FR-15K and SRPRS_EN-DE-15K are cross-lingual datasets, and SRPRS_D-Y-15K is a monolingual dataset. The detailed information of these datasets is shown in Table 1.

Table 1. General-domain datasets.

Dataset	KG Source	Entities	Relations	Attributes	Rel. Triples	Attr. Triples	Aligned Entities
SRPRS_EN-FR-15K	DBpedia_English DBpedia_French	15,000 15,000	267 210	308 404	47,334 40,864	73,121 67,167	15,000
SRPRS_EN-DE-15K	DBpedia_English DBpedia_German	15,000 15,000	215 131	286 194	47,676 50,419	83,755 156,150	15,000
SRPRS_D-Y-15K	DBpedia YAGO	15,000 15,000	165 128	257 135	30,291 26,638	71,716 132,114	15,000

Pipe-DPMEA is an entity alignment dataset for the oil and gas pipeline safety domain constructed using data provided by a Chinese oil and gas pipeline company. The dataset language is Chinese (monolingual). It primarily collects various types of risk events and their associated oil and gas pipeline safety data, such as preventive measures, hazard sources, and mitigation measures, forming an entity alignment dataset for the oil and gas pipeline safety domain containing two knowledge graphs to be aligned, providing fundamental data support for entity alignment tasks in oil and gas pipeline safety. The detailed information of this dataset is shown in Table 2.

Table 2. Oil and gas pipeline safety domain entity alignment dataset.

Dataset	Entities	Relations	Attributes	Rel. Triples	Attr. Triples	Aligned Entities
Pipe-DPMEA	482 431	4 4	4 3	529 466	1493 1293	311

5.2. Implementation Details

The hardware platform employed for all experiments comprises an Nvidia GeForce RTX 4060 Ti GPU with 16 GB of memory (Nvidia Corporation, Santa Clara, CA, USA),

paired with an Intel i5-12600KF CPU (Intel Corporation, Santa Clara, CA, USA). The parameters for the semantic name matrix α_1, semantic attribute matrix α_2, and edit distance matrix α_3 are set to 0.6, 0.3, and 0.1, respectively. The seed ratio is $\beta = 0.2$, the time smoothing coefficient is $\mu = 2/3$, $\lambda_c = 0.1$, the semantic weight coefficient is $\delta = 0.7$, the BFS maximum depth is $d_{\max} = 2$, and the similarity threshold (τ) is initialized to $\tau_{\text{high}} = 0.95$ and decreases by 0.01 every 10 epochs after 30 training epochs until it reaches $\tau_{\text{low}} = 0.8$. For the encoder, the number of network layers is layer $= 4$, the embedding vector dimension per layer is dim $= 256$, the learning rate is lr $= 0.0001$, and the Adam optimizer is used with a relation update ratio of $\varepsilon = 0.8$. For each pre-aligned entity pair, 20 negative samples are randomly sampled. The batch size is set to 256 for the oil and gas pipeline safety entity alignment dataset due to the smaller number of entities and 1024 for all other datasets. All methods were implemented in Python 3.9.22 with PyTorch 2.6.0 (CUDA 11.8) and PyTorch Geometric 2.6.1.

For a fair comparison, each supervised baseline is trained on a random 30% split of the ground-truth alignment pairs, while the remaining 70% serve as the test set; unsupervised models are evaluated directly on the same 70% test partition. All reported metrics are averaged across five independent runs. Experiments on the SRPRS_EN-FR-15K and SRPRS_EN-DE-15K datasets do not use machine translation. The experimental results of CAEA, RREA, and BootEA on the publicly available general-domain datasets are cited from [19].

To accurately evaluate model performance and conduct benchmark comparisons, we adopt Hits@k ($k = 1, 10$) and Mean Reciprocal Rank (MRR), which are commonly used evaluation metrics in entity alignment. Hits@k reflects the proportion of correct aligned entities within the top k ranked entities, where Hits@1 is the accuracy of entity alignment. MRR computes the mean of the reciprocal ranks of correct entities, reflecting the model's ranking ability for target entities from a global perspective.

5.3. Baselines

To benchmark the proposed method against existing approaches, nine representative entity alignment techniques are adopted as baselines, falling into two broad categories. The first class relies solely on structural information (*structure-based*), including BootEA, AliNet, Dual-AMN, RHGN, and MRAEA. The second class leverages not only structural information but also semantic information, such as attributes and entity names (*semantic-enhanced*), including RREA, AttrGNN, CAEA, and SelfKG.

5.4. Experimental Results

The experimental results of MIF-UEA on the Pipe-DPMEA dataset are shown in Table 3; the SelfKG model does not provide the MRR evaluation metric, which is marked with "–" in the table. MIF-UEA achieves the best performance across all evaluation metrics, with Hits@1, Hits@10, and MRR reaching 0.9871, 1.0000, and 0.994, respectively. Compared to the second-best SelfKG model, MIF-UEA improves the core Hits@1 metric by 3.81%, fully demonstrating the significant advantages of this model in handling knowledge graphs in the oil and gas pipeline safety domain.

Models that rely solely on graph structure (e.g., AliNet and BootEA) perform poorly on this domain-specific dataset, indicating that knowledge graphs in the oil and gas pipeline domain suffer from severe structural sparsity. In contrast, semantic-enhanced models that leverage attribute and name information (e.g., AttrGNN, CAEA) achieve significant performance improvements.

The experimental results of the proposed MIF-UEA model on the SRPRS_EN-FR-15K, SRPRS_EN-DE-15K, and SRPRS_D-Y-15K datasets are shown in Table 4.

Table 3. Experimental results on the Pipe-DPMEA dataset.

Model	Hits@1	Hits@10	MRR
Structure-based methods			
AliNet	0.3073	0.7477	0.446
BootEA	0.3578	0.6835	0.452
MRAEA	0.3748	0.8413	0.521
RHGN	0.4119	0.9353	0.557
Dual-AMN	0.4431	0.8330	0.567
Semantically enhanced methods			
AttrGNN	0.7661	0.9495	0.839
CAEA	0.8633	0.9578	0.898
RREA	0.9037	0.9908	0.940
SelfKG	0.9490	0.9970	–
MIF-UEA (Ours)	**0.9871**	**1.0000**	**0.994**

Table 4. Experimental results on general-domain datasets.

Model	SRPRS_EN-FR-15K			SRPRS_EN-DE-15K			SRPRS_D-Y-15K		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
Structure-based methods									
AliNet	0.4760	0.8065	0.589	0.6661	0.8890	0.739	0.6910	0.8480	0.751
BootEA	0.5031	0.7861	0.597	0.6714	0.8660	0.767	0.7387	0.8707	0.786
MRAEA	0.6075	0.8869	0.706	0.7386	0.9346	0.805	0.7867	0.9220	0.839
Dual-AMN	0.6913	0.9267	0.775	0.8032	0.9571	0.858	0.8423	0.9318	0.879
RHGN	0.7315	0.9291	0.799	0.5738	0.8824	0.680	0.8431	0.9443	0.882
Semantically enhanced methods									
RREA	0.8327	0.9508	0.877	0.9005	0.9725	0.927	0.9517	0.9785	0.961
AttrGNN	0.8868	0.9515	0.911	0.9263	0.9741	0.945	0.9676	0.9837	0.974
CAEA	0.9240	0.9595	0.937	0.9323	0.9797	0.951	**0.9999**	**1.0000**	**1.000**
SelfKG	0.9260	0.9760	–	0.9340	0.9260	–	0.9980	1.0000	–
MIF-UEA (Ours)	**0.9913**	**0.9980**	**0.994**	**0.9935**	**0.9985**	**0.995**	0.9993	1.0000	1.000

The experimental results demonstrate that MIF-UEA exhibits excellent performance in graph alignment tasks across different scales and language settings. In the two cross-lingual scenarios of SRPRS_EN-FR-15K and SRPRS_EN-DE-15K, the Hits@1 of MIF-UEA reaches 0.9913 and 0.9935, respectively, outperforming all comparison models, including CAEA and SelfKG. Notably, MIF-UEA achieves these results under a completely unsupervised setting, while CAEA, RREA, and other models rely on 30% labeled alignment pairs as training signals, further highlighting the advantages of the proposed method in cross-lingual entity alignment tasks. On the monolingual SRPRS_D-Y-15K dataset, the Hits@1 of MIF-UEA reaches 0.9993, which is slightly lower than CAEA's 0.9999. The analysis suggests that in this dataset, since the source and target graphs share the same language and possess relatively complete topological structures, entity names and attributes alone already provide extremely strong alignment signals. Most models that incorporate semantic information can achieve near-perfect performance in this scenario, and the performance differences among models are negligible. A one-sample t-test over five independent runs of MIF-UEA (mean Hits@1 = 0.9993, std = 0.00057) against CAEA's reported value of 0.9999 yields $t = -2.353$, $p = 0.078$, confirming that the difference is not statistically significant at the 0.05 level. Notably, this comparison is between CAEA (using 30% labeled alignment pairs) and MIF-UEA (fully unsupervised). In contrast, the advantages of MIF-UEA are more fully demonstrated in cross-lingual scenarios and domain scenarios with sparse structures.

5.5. Ablation Study

To investigate the effectiveness of each component in the MIF-UEA model for entity alignment tasks in the oil and gas pipeline safety domain, ablation experiments are conducted on the Pipe-DPMEA dataset, comparing the following four variants: **MIF-UEA w/o StructureSeed** removes the structure-aware multi-source similarity seed generation algorithm (Algorithm 1), relying solely on the original similarity matrix S_{final} for initial matching. **MIF-UEA w/o Semi** removes the pseudo-label augmentation and removal module, and the model only uses initial seeds for training, without performing iterative pseudo-label mining and dynamic denoising. **MIF-UEA w/o Attribute** removes the semantic attribute representation learning module, where the model relies solely on the structural information extracted by the Relation-Gated Convolution (RGC) network for alignment. **MIF-UEA w/o RGC** removes the structural representation learning module (RGC), and the model relies solely on attribute and name semantic information extracted by the pre-trained language model and attention mechanism for alignment. The experimental results are shown in Table 5.

Table 5. Experimental ablation results on the Pipe-DPMEA dataset.

Model	Hits@1	Hits@10	MRR
MIF-UEA w/o Attribute	0.8039	0.9743	0.860
MIF-UEA w/o RGC	0.9518	1.0000	0.975
MIF-UEA w/o Semi	0.9646	1.0000	0.981
MIF-UEA w/o StructureSeed	0.9775	1.0000	0.989
MIF-UEA	**0.9871**	**1.0000**	**0.994**

At the feature-encoding level, removing attribute features (MIF-UEA w/o Attribute) leads to the most significant performance degradation, with Hits@1 dropping sharply from 98.71% to 80.39%. This is primarily because knowledge graphs in the oil and gas pipeline safety domain generally suffer from structural sparsity, and in the absence of attribute descriptions, it is extremely difficult to distinguish similar entities relying solely on sparse connectivity structures. Removing structural features (MIF-UEA w/o RGC) also results in Hits@1 decreasing to 95.18%. The complete framework yields the best results across all three evaluation criteria, offering compelling proof that attribute semantics and graph structural topology are highly complementary within the deep representation space. Information from a single modality is insufficient to support the alignment of complex domain knowledge graphs, and multi-information fusion is the key to improving the accuracy of entity alignment in this domain.

Comparing the complete model with MIF-UEA w/o StructureSeed reveals that removing the structure-aware mechanism leads to a 0.96% decrease in Hits@1. Without the BFS-based structural penalty mechanism, the initial alignment seeds tend to be excessively concentrated in locally high-density regions of the graph, making it difficult for the model to acquire recognition capability for entities in globally sparse regions. This mechanism effectively ensures a more balanced distribution of prior seeds across the graph space, thereby improving the global quality of initial supervision signals. Furthermore, comparing the complete model with MIF-UEA w/o Semi shows that removing the pseudo-label augmentation and denoising module results in a 2.25% decrease in Hits@1. This is because the number of initial seed pairs under unsupervised conditions is relatively limited, and relying solely on initialization information is insufficient to support precise learning of deep cross-graph mapping relationships. Through the high-confidence pseudo-label augmentation and exponential moving average denoising strategies, reliable sample pairs are

iteratively expanded, effectively mitigating error accumulation and significantly improving the model's final performance.

To further validate the generalizability of these findings, ablation experiments are conducted on three general-domain datasets. The results are presented in Table 6.

Table 6. Experimental ablation results on general-domain datasets.

Model	SRPRS_EN-FR-15K			SRPRS_EN-DE-15K			SRPRS_D-Y-15K		
	Hits@1	**Hits@10**	**MRR**	**Hits@1**	**Hits@10**	**MRR**	**Hits@1**	**Hits@10**	**MRR**
w/o Attribute	0.8962	0.9797	0.928	0.9190	0.9887	0.945	0.9684	0.9985	0.980
w/o RGC	0.9797	0.9886	0.983	0.9447	0.9745	0.955	0.9991	1.0000	1.000
w/o Semi	0.9723	0.9958	0.981	0.9783	0.9958	0.985	0.9929	0.9990	0.995
w/o StructureSeed	0.9879	0.9975	0.991	0.9880	0.9972	0.993	0.9979	0.9995	0.998
MIF-UEA	**0.9913**	**0.9980**	**0.994**	**0.9935**	**0.9985**	**0.995**	**0.9993**	**1.0000**	**1.000**

The ablation results on general-domain datasets confirm that each module contributes meaningfully across different scenarios. Removing attribute features (w/o Attribute) also causes notable performance drops on general-domain datasets, though less dramatic than on Pipe-DPMEA, likely because general-domain graphs possess richer structural connectivity that can partially compensate for the absence of attribute information. The impact of removing structural features (w/o RGC) varies across datasets: the performance drop is negligible on the monolingual SRPRS_D-Y-15K dataset, whereas it reaches 4.88 percentage points on the cross-lingual SRPRS_EN-DE-15K dataset, indicating that structural features play a more critical role in bridging the semantic gap when language differences are larger. The pseudo-label augmentation and denoising module (w/o Semi) demonstrates consistent importance across all datasets. The structure-aware seed generation mechanism (w/o StructureSeed) yields relatively smaller improvements on general-domain datasets, possibly because similarity-based matching on larger and denser graphs already tends to produce reasonably well-distributed seeds, limiting the marginal benefit of the structure-aware penalty mechanism.

5.6. Parameter Sensitivity Analysis of Semantic Similarity Matrix

To examine how varying the weight coefficients assigned to the semantic name matrix, the semantic attribute matrix, and the edit distance matrix influences alignment outcomes, a series of controlled experiments is carried out on the Pipe-DPMEA dataset. The α_1, α_2, and α_3 parameters represent the weight parameters of the semantic name matrix, semantic attribute matrix, and edit distance matrix, respectively. The experimental results and settings are shown in Table 7.

Table 7. Experimental results of semantic similarity matrix weights.

Parameter Setting	Hits@1	Hits@10	MRR
$\alpha_1 = 1, \alpha_2 = 0, \alpha_3 = 0$	0.9742	1.0000	0.986
$\alpha_1 = 0, \alpha_2 = 1, \alpha_3 = 0$	0.9679	1.0000	0.981
$\alpha_1 = 0, \alpha_2 = 0, \alpha_3 = 1$	0.9711	1.0000	0.983
$\alpha_1 = 0.5, \alpha_2 = 0.5, \alpha_3 = 0$	0.9774	1.0000	0.988
$\alpha_1 = 0.5, \alpha_2 = 0, \alpha_3 = 0.5$	0.9679	1.0000	0.981
$\alpha_1 = 0, \alpha_2 = 0.5, \alpha_3 = 0.5$	0.9581	1.0000	0.979
$\alpha_1 = 0.6, \alpha_2 = 0.3, \alpha_3 = 0.1$	**0.9871**	**1.0000**	**0.994**

When the model relies on a single feature for alignment, the semantic name matrix(α_1) achieves the best single-feature performance, with Hits@1 reaching 97.42% and MRR

reaching 0.986, indicating that entity names inherently contain the most direct and important alignment cues. When only the semantic attribute matrixor edit distance matrix is used, the performance is slightly lower than that using the semantic name matrix, but Hits@1 can still reach 96.79% and 97.11%, respectively. When two features are assigned equal weights for fusion, the combination of name semantics and attribute semantics ($\alpha_1 = 0.5, \alpha_2 = 0.5, \alpha_3 = 0$) improves Hits@1 to 97.74% and MRR to 0.988, outperforming the individual effects of either feature. This demonstrates that name and attribute information exhibit good complementarity in the deep semantic space. However, when name or attribute features are fused with edit distance at equal weights ($\alpha_1 = 0.5, \alpha_2 = 0, \alpha_3 = 0.5$ and $\alpha_1 = 0, \alpha_2 = 0.5, \alpha_3 = 0.5$), the alignment metrics actually decrease to varying degrees. This indicates that directly fusing name or attribute information with edit distance at equal weights causes mutual interference between the information sources, resulting in declining rather than improving alignment metrics. By assigning different weights to the three types of matrices ($\alpha_1 = 0.6, \alpha_2 = 0.3, \alpha_3 = 0.1$), Hits@1 reaches 98.71% and MRR reaches 0.994, exhibiting optimal performance. This result demonstrates that a weighted strategy with name semantics as the primary component, attribute semantics as a secondary component, and edit distance as a supplement can fully leverage the complementary advantages of various information types to achieve optimal entity alignment performance.

5.7. Parameter Sensitivity Analysis of Similarity Threshold

In the unsupervised pseudo-label augmentation strategy of the MIF-UEA framework, the similarity threshold (τ) plays a crucial role. Based on the variation bounds, it can be divided into the initial high-confidence threshold (τ_{high}) and the annealing lower-bound threshold (τ_{low}). The initial high-confidence threshold (τ_{high}) determines the purity of the prior knowledge acquired by the model, while the annealing lower-bound threshold (τ_{low}) determines the lower limit for mining hard samples during the middle and late stages of self-training. To thoroughly investigate these two key hyperparameters, two-dimensional grid-search experiments are conducted on the Pipe-DPMEA dataset. The experimental results are shown in Figure 3.

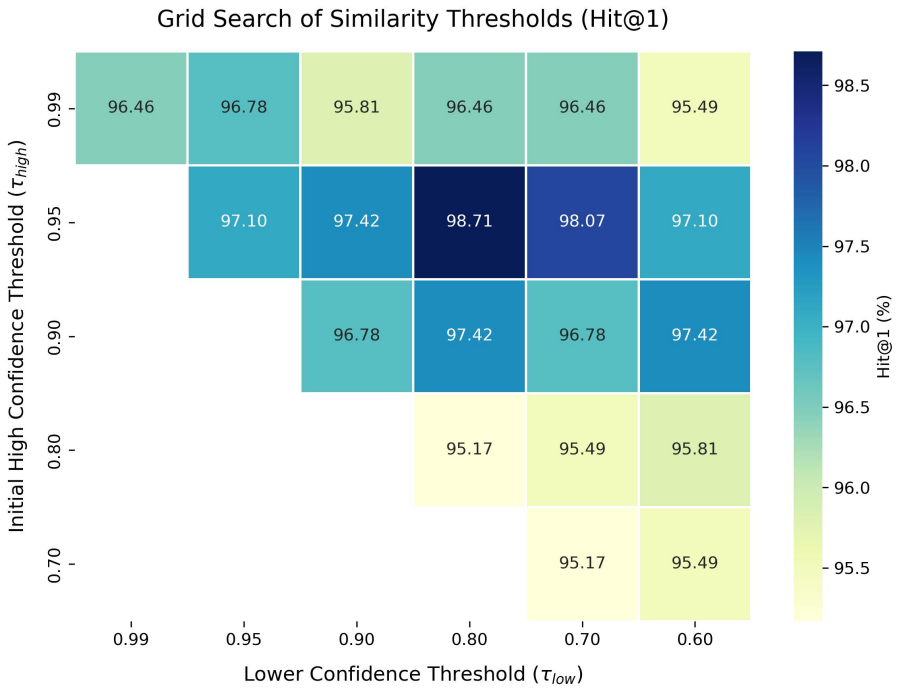

Figure 3. Experimental results of similarity threshold parameters.

The results reveal that the model's alignment accuracy is highly sensitive to the choice of threshold settings, achieving global optimum at the combination of $\tau_{high} = 0.95$ and $\tau_{low} = 0.8$. Specifically, the initial threshold (τ_{high}) determines the quality and quantity of initial supervision signals. Setting it too high leads to an excessively small initial seed set, causing the graph neural network to suffer from underfitting due to insufficient supervision signals. Setting it too low introduces a large amount of noise in the early stages of self-training, directly damaging the model's feature representation. Meanwhile, the annealing lower bound (τ_{low}) determines the lower limit for mining hard samples. When $\tau_{low} = 0.8$, the model can maintain optimal performance under most parameter combinations. However, when it is further reduced to 0.6, the alignment accuracy of all configurations decreases to varying degrees because the model is forced to incorporate a large number of completely unrelated noise entities in the later iterations, leading to performance degradation.

5.8. Impact of GNN Layer Depth

To explore how the depth of GNN layers influences the accuracy of entity alignment, we perform evaluations using the Pipe-DPMEA benchmark. Specifically, we adjust the Relation-Gated Convolution (RGC) layer count from 1 to 6 while holding all other configurations constant. The experimental results are shown in Figure 4.

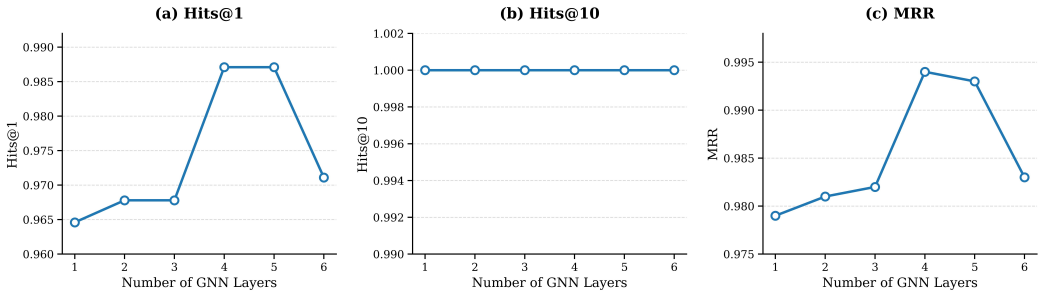

Figure 4. Impact of GNN layer depth on entity alignment performance on the Pipe-DPMEA dataset.

The results reveal a clear trend of initial improvement followed by degradation as the network depth increases. With a single layer, the model can only aggregate information from immediate neighbors, limiting its ability to capture higher-order structural patterns. Increasing the depth to two and three layers yields modest improvements, as the model begins to leverage multi-hop neighborhood information. A notable performance jump occurs at four layers, where the model achieves optimal results on all three metrics, indicating that this depth provides a sufficient receptive field to capture the meaningful structural context in the Pipe-DPMEA knowledge graph. At five layers, Hits@1 remains unchanged while MRR slightly decreases, suggesting that the additional layer provides no further benefit. When the depth is further increased to six layers, both Hits@1 and MRR drop noticeably, as excessively deep aggregation causes entity representations to converge toward indistinguishable embeddings, thereby reducing the discriminative power of the learned features. Given that the Pipe-DPMEA dataset is characterized by structural sparsity, with only a few neighbors per entity, on average, a four-layer network achieves the optimal balance between capturing sufficient neighborhood context and avoiding over-smoothing.

5.9. Impact of Pre-Trained Language Model

To evaluate how various pre-trained language models affect alignment accuracy, we test our approach on the Pipe-DPMEA dataset utilizing BERT and LaBSE as alternative text encoders. The corresponding outcomes are illustrated in Figure 5.

The results show that LaBSE outperforms BERT on all three evaluation metrics, with Hits@1 improving by 2.25 percentage points and MRR improving by 1.6 percentage points. Notably, even with the BERT encoder, the model still achieves a Hits@1 of 96.46%, indicating that the MIF-UEA framework exhibits a certain degree of robustness to the choice of encoder, as structural information and the self-training mechanism can partially compensate for weaker semantic encoding. However, the notable improvements in Hits@1 and MRR with LaBSE suggest that higher-quality semantic representations provide more reliable initial signals for seed generation and pseudo-label selection, producing a cumulative positive effect throughout the self-training process. These results validate the choice of LaBSE as the text encoder in the proposed method.

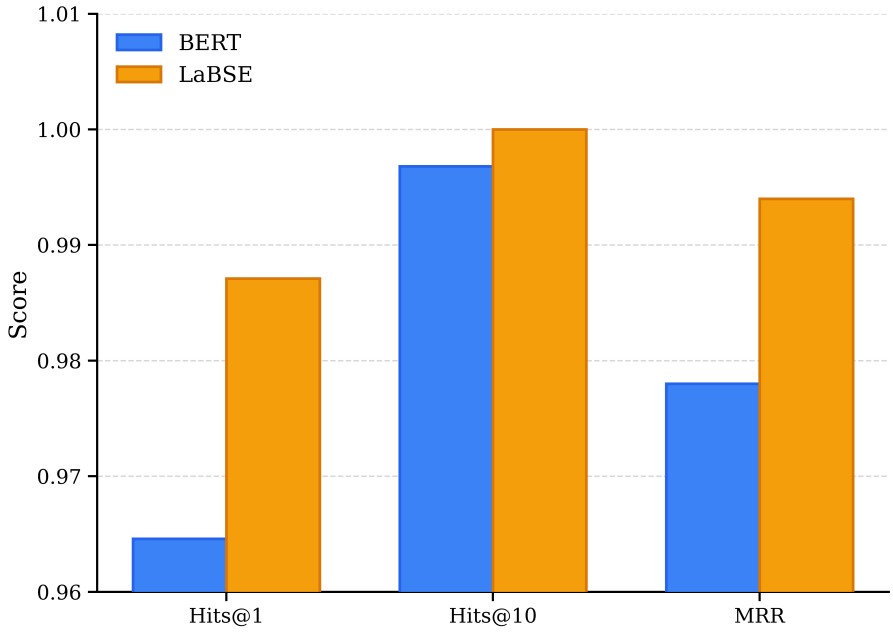

Figure 5. Comparison of entity alignment performance using different pre-trained language models on the Pipe-DPMEA dataset.

6. Conclusions

This paper addresses the challenges of structural sparsity, insufficient semantic information, and lack of labeled data in knowledge graphs for the oil and gas pipeline safety domain and proposes a Multi-Information Fusion Unsupervised Entity Alignment model (MIF-UEA). The proposed method achieves high-accuracy entity alignment without manual annotations through structure–semantic fusion representation learning, high-quality seed generation based on multi-source information, self-training pseudo-label augmentation and denoising, and an optimal transport-based matching strategy. Empirical evaluations show that MIF-UEA consistently outperforms competing approaches on both the domain-specific oil and gas pipeline safety dataset and several general-purpose benchmarks, substantiating the effectiveness and generalizability of the proposed framework. Moreover, systematic ablation studies verify that every constituent module makes a meaningful contribution to the overall alignment quality.

Nevertheless, this study has certain limitations. The Pipe-DPMEA dataset is relatively small in scale, and the generalizability of the proposed method to larger and more diverse industrial knowledge graphs remains to be further validated. Additionally, the current approach relies on a static knowledge graph structure and does not account for temporal changes in entity relationships. Looking ahead, several promising directions merit further investigation: (1) incorporating large language models to strengthen the quality of semantic

representations; (2) devising scalable alignment strategies tailored to large-scale, temporally evolving knowledge graphs; (3) adapting the proposed framework to broader industrial safety scenarios beyond the oil and gas pipeline domain; and (4) collaborating with more pipeline operators to construct larger-scale oil and gas pipeline safety entity alignment datasets, thereby further validating the scalability of the proposed method.

Author Contributions: Conceptualization, K.L. and W.C.; methodology, H.D.; software, W.S. and H.D.; validation, W.S., H.D. and G.Z.; formal analysis, W.S. and H.D.; investigation, W.S., H.D. and G.Z.; resources, W.C.; data curation, H.D. and W.C.; writing—original draft preparation, H.D.; writing—review and editing, W.S., K.L. and G.Z.; visualization, W.S. and H.D.; supervision, K.L. and W.C.; project administration, K.L.; funding acquisition, K.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the major project of the National Natural Science Foundation of China (51991365) and the Natural Science Foundation of Shandong Province of China (ZR2021MF082).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The publicly available general-domain datasets (SRPRS_EN-FR-15K, SRPRS_EN-DE-15K, and SRPRS_D-Y-15K) used in this study are openly available at https://github.com/ws-researcher/CAEA (accessed on 23 March 2026). The Pipe-DPMEA dataset that supports the findings of this study is available from the corresponding author upon reasonable request.

Conflicts of Interest: Author Weichun Chang was employed by PipeChina Science and Technology Research Institute. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

KG	Knowledge Graph
EA	Entity Alignment
GNN	Graph Neural Network
MIF-UEA	Multi-Information Fusion Unsupervised Entity Alignment
RGC	Relation-Gated Convolution
CSLS	Cross-domain Similarity Local Scaling
BFS	Breadth-First Search
MRR	Mean Reciprocal Rank
LaBSE	Language-agnostic BERT Sentence Embedding
EMA	Exponential Moving Average

References

1. Wu, Z. Construction and application of knowledge graph for geological disaster risk management of oil and gas pipelines. *Oil Gas Storage Transp.* **2023**, *42*, 241–248.
2. Chen, C.; Hu, J.; Han, Z.; Chen, Y.; Xiao, S. Knowledge graph modeling and early warning method for accident evolution of overseas natural gas pipeline stations under harsh environmental conditions. *J. Tsinghua Univ. (Sci. Technol.)* **2022**, *62*, 1081–1087.
3. Zhao, X.; Zeng, W.; Tang, J.; Wang, W.; Suchanek, F. An experimental study of state-of-the-art entity alignment approaches. *IEEE Trans. Knowl. Data Eng.* **2020**, *34*, 2610–2625.
4. Zeng, K.; Li, C.; Hou, L.; Li, J.; Feng, L. A comprehensive survey of entity alignment for knowledge graphs. *AI Open* **2021**, *2*, 1–13. [CrossRef]
5. 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.
6. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating embeddings for modeling multi-relational data. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 2787–2795.

7. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4–24. [CrossRef]
8. Khemani, B.; Patil, S.; Kotecha, K.; Tanwar, S. A review of graph neural networks: Concepts, architectures, techniques, challenges, datasets, applications, and future directions. *J. Big Data* **2024**, *11*, 18. [CrossRef]
9. Chen, M.; Tian, Y.; Yang, M.; Zaniolo, C. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. *arXiv* **2016**, arXiv:1611.03954.
10. Sun, Z.; Hu, W.; Zhang, Q.; Qu, Y. Bootstrapping entity alignment with knowledge graph embedding. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 4396–4402.
11. Sun, Z.; Wang, C.; Hu, W.; Chen, M.; Dai, J.; Zhang, W.; Qu, Y. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 222–229.
12. Mao, X.; Wang, W.; Xu, H.; Lan, M.; Wu, Y. MRAEA: An efficient and robust entity alignment approach for cross-lingual knowledge graph. In Proceedings of the 13th International Conference on Web Search and Data Mining, Houston, TX, USA, 3–7 February 2020; pp. 420–428.
13. 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 Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 821–832.
14. Liu, X.; Zhang, K.; Liu, Y.; Chen, E.; Huang, Z.; Yue, L.; Yan, J. RHGN: Relation-gated heterogeneous graph network for entity alignment in knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2023*; Association for Computational Linguistics: Toronto, ON, Canada, 2023; pp. 8683–8696.
15. Zhang, H.; Zhao, H.; Fu, Y.; Ma, J.; Xiang, Y. The class labels and spatial information based fault diagnosis of air handling unit via combining kernel Fischer discriminant analysis with an improved graph convolutional neural network. *Measurement* **2026**, *257*, 118622.
16. Zhang, Q.; Sun, Z.; Hu, W.; Chen, M.; Guo, L.; Qu, Y. Multi-view knowledge graph embedding for entity alignment. *arXiv* **2019**, arXiv:1906.02390.
17. Liu, Z.; Cao, Y.; Pan, L.; Li, J.; Liu, Z.; Chua, T.S. Exploring and evaluating attributes, values, and structures for entity alignment. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 6355–6364.
18. Mao, X.; Wang, W.; Xu, H.; Wu, Y.; Lan, M. Relational reflection entity alignment. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual Event, Galway, Ireland, 19–23 October 2020; pp. 1095–1104.
19. Wu, S.; Tong, W.; Hou, Y.; Li, P. Concept-Aware Entity Alignment Network for Industrial Knowledge Graph. *IEEE Trans. Ind. Inform.* **2025**, *21*, 4316–4323. [CrossRef]
20. Liu, X.; Hong, H.; Wang, X.; Chen, Z.; Kharlamov, E.; Dong, Y.; Tang, J. SelfKG: Self-supervised entity alignment in knowledge graphs. In Proceedings of the ACM Web Conference 2022, Virtual Event, Lyon, France, 25–29 April 2022; pp. 860–870.
21. Jiang, C.; Qian, Y.; Chen, L.; Gu, Y.; Xie, X. Unsupervised deep cross-language entity alignment. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Turin, Italy, 18–22 September 2023*; Springer: Cham, Switzerland, 2023; pp. 3–19.
22. Bai, Y.; Wu, J.; Ren, Q.; Jiang, Y.; Cai, J. A BN-based risk assessment model of natural gas pipelines integrating knowledge graph and DEMATEL. *Process Saf. Environ. Prot.* **2023**, *171*, 640–654. [CrossRef]
23. Chen, Y.; Zhang, L.; Hu, J.; Chen, C.; Fan, X.; Li, X. An emergency task recommendation model of long-distance oil and gas pipeline based on knowledge graph convolution network. *Process Saf. Environ. Prot.* **2022**, *167*, 651–661. [CrossRef]
24. Simone, F.; Ansaldi, S.M.; Agnello, P.; Patriarca, R. Industrial safety management in the digital era: Constructing a knowledge graph from near misses. *Comput. Ind.* **2023**, *146*, 103849. [CrossRef]
25. Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; Wang, W. Language-agnostic BERT sentence embedding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 878–891.
26. Yujian, L.; Bo, L. A normalized Levenshtein distance metric. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1091–1095. [CrossRef] [PubMed]
27. Conneau, A.; Lample, G.; Ranzato, M.A.; Denoyer, L.; Jégou, H. Word translation without parallel data. *arXiv* **2017**, arXiv:1710.04087.
28. Wu, Y.; Liu, X.; Feng, Y.; Wang, Z.; Zhao, D. Jointly learning entity and relation representations for entity alignment. *arXiv* **2019**, arXiv:1909.09317.
29. Zhu, Y.; Liu, H.; Wu, Z.; Du, Y. Relation-aware neighborhood matching model for entity alignment. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; Volume 35, pp. 4749–4756.
30. Vashishth, S.; Sanyal, S.; Nitin, V.; Talukdar, P. Composition-based multi-relational graph convolutional networks. *arXiv* **2019**, arXiv:1911.03082.

31. Gui, J.; Chen, T.; Zhang, J.; Cao, Q.; Sun, Z.; Luo, H.; Tao, D. A survey on self-supervised learning: Algorithms, applications, and future trends. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 9052–9071. [CrossRef] [PubMed]
32. Luo, S.; Yu, S. An accurate unsupervised method for joint entity alignment and dangling entity detection. In *Findings of the Association for Computational Linguistics: ACL 2022*; Association for Computational Linguistics: Dublin, Ireland, 2022; pp. 2330–2339.
33. Mao, X.; Wang, W.; Wu, Y.; Lan, M. From alignment to assignment: Frustratingly simple unsupervised entity alignment. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 2843–2853.
34. Tang, J.; Zhao, K.; Li, J. A fused Gromov-Wasserstein framework for unsupervised knowledge graph entity alignment. In *Findings of the Association for Computational Linguistics: ACL 2023*; Association for Computational Linguistics: Toronto, ON, Canada, 2023.
35. Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 2292–2300.