EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

MFSF-CEA: A Multi-Feature Similarity Fusion Model for Chinese Entity Alignment

Min Zhang*, Luya Yang, Yaxian Gao, Lina Han and Bai Caimei

School of Information Engineering, Shaanxi Xueqian Normal University, Xi'an, Shaanxi, 710010, China

Abstract

INTRODUCTION: Entity alignment across multi-source encyclopedic knowledge bases is crucial for constructing high-quality knowledge graphs. This task is particularly challenging in specialized vertical domains like Chinese cultural relics, where heterogeneous data sources and diverse descriptive patterns render single-feature alignment methods inadequate.
OBJECTIVES: To address this challenge, we propose MFSF-CEA, a Multi-feature Similarity Fusion model featuring dual-layer optimization: multi-granularity semantic modeling and domain-adaptive dynamic weight fusion.
METHODS: This approach employs a three-tiered semantic capture structure: character-level similarity using Longest Common Subsequence for variant character matching; word-level similarity via TF-IDF for core concept association; and sentence-level semantic similarity through Latent Dirichlet Allocation for deep topic alignment. Beyond feature extraction, we introduce an Entropy-AHP combined weighting mechanism that dynamically balances objective information contribution and domain expert knowledge, overcoming limitations of fixed-weight fusion strategies. Experimental evaluation on a Chinese cultural relics dataset demonstrates that MFSF-CEA significantly outperforms baseline methods in precision, recall, and F1-score. The sentence-level contextual features contribute most substantially to alignment accuracy, while the multi-feature fusion effectively compensates for the limitations of any single feature type, particularly the sparsity of word-level abstract features.
RESULTS: The proposed framework successfully addresses the unique challenges of entity alignment in cultural relic texts by leveraging complementary features across multiple linguistic levels.
CONCLUSION: This work provides an effective and extensible solution for knowledge fusion in vertical domains, advancing entity alignment from traditional string matching toward deeper semantic integration.

Keywords: Entity alignment, multi-feature, entity attribute, entity abstract, LDA

Received on 29 October 2025, accepted on 24 March 2025, published on 08 April 2026

doi: 10.4108/eetsis.10735

1. Introduction

The exponential growth of structured and semi-structured knowledge sources has made knowledge graphs (KGs) a cornerstone for many AI applications. After knowledge extraction, a considerable amount of formal knowledge can be assembled. However, the inherent diversity and heterogeneity of these sources such as Baidu Baikie and Chinese Wikipedia often lead to significant conflicts and overlaps among knowledge items [1, 2]. Consequently, an initially constructed KG is typically insufficient in both quality and coverage, presenting a major bottleneck for downstream tasks [3,4]. Entity alignment (EA) aims to identify and link semantically equivalent entities that refer to the same real-world object across different KGs. It is thus a critical step in the knowledge fusion process. [5].

Current research on entity alignment for encyclopaedic knowledge bases has given rise to several typical technical approaches [6,7]. The majority of these methods heavily depend on the structural or semantic similarity measures

* Corresponding author. Email:28028@snsy.edu.cn

inherent within the knowledge graph systems themselves [8,9]. Furthermore, another prominent class of techniques focuses on extracting and utilizing multi-faceted information from entities themselves, such as their attributes and names, to compute similarities across different knowledge sources for alignment purposes [10]. Nevertheless, when dealing with large-scale, highly heterogeneous real-world knowledge graphs, these conventional methods still face significant challenges in terms of their accuracy and robustness. For example, Simos et al. [1] employed methods such as fuzzy logic and relative commonness to achieve efficient context-free Vilification, while Szymański et al. [2] utilized contextual information of entities for semantic disambiguation, thereby improving linking accuracy. However, these approaches predominantly depend on the built-in disambiguation mechanisms of encyclopaedic sources, utilize only single-dimensional entity features, and fail to fully leverage the potential of integrating multiple types of knowledge available in encyclopaedic repositories.

To overcome the constraints of single-feature alignment, several studies have explored multi-feature fusion strategies. Bai et al. [11] proposed a joint multi-feature entity alignment method for cross-lingual temporal knowledge graphs, which improves entity alignment performance by learning temporal information embeddings. Song et al. [12] proposed an ontology-enhanced entity alignment method based on multi-feature fusion, which leverages different semantic features to achieve more accurate alignment. While these methods represent advancements through feature combination, their similarity calculations for entity attributes and abstracts remain largely confined to the character or word level, essentially constituting shallow string-based comparisons.

Two critical research challenges persist: first, they fail to capture deeper semantic relationships and contextual associations at the sentence or discourse level, resulting in insufficient recognition of complex scenarios where expressions differ but meanings converge—such as synonyms or paraphrases [13]; second, they overlook "feature importance disparities" in vertical domains. For example, in the field of Chinese cultural relics, attributes like "cultural relic age" and "excavation site" contribute far more to alignment than "relic size" or "collecting institution."[14,15] However, existing multi-feature fusion methods mostly adopt equal weights or empirical weights, which cannot dynamically adapt to domain-specific characteristics. This often leads to the downplaying of key features and interference from redundant features in alignment outcomes.

These limitations are particularly prominent in vertical domains with specialized terminologies and knowledge structures—such as Chinese cultural relics. Descriptions in this field typically involve professional terminology, historical contexts, and nuanced semantic relationships, which shallow matching techniques struggle to handle. Furthermore, the significant variability in feature discriminability within the domain, combined with fixed-weight fusion approaches, further degrades alignment accuracy.

To address the aforementioned challenges, this paper proposes a multi-feature similarity fusion-based entity alignment method for Chinese cultural relic texts, aiming to overcome the limitations of single-feature alignment and improve the accuracy of cross-encyclopaedic knowledge graph fusion. The core innovation of this study lies in the construction of a dual-layer optimization framework encompassing "multi-granularity semantic capture" and "dynamic weight fusion," which overcomes existing limitations in both feature representation and fusion strategies. The framework is detailed as follows.

(1) Innovation in Multi-Granularity Semantic Modelling: A three-tier feature fusion structure is designed to comprehensively capture entity semantics across different dimensions:

- At the character level, the Longest Common Subsequence (LCS) algorithm is used to calculate fine-grained similarity for entity attributes, enabling accurate matching of "variant character expressions".
- At the word level, Term Frequency-Inverse Document Frequency (TF-IDF) is employed to extract keywords from entity abstracts and measure their similarity, focusing on the association of core concepts.
- At the sentence level, the Latent Dirichlet Allocation (LDA) model is applied to extract latent semantic topics from the broader contextual information of entities. This facilitates the identification of deep-level associations where "expressions differ but topics align".

(2)Innovation in Domain-Adaptive Dynamic Weight Fusion: To address feature importance disparities in the Chinese cultural relic domain, an Entropy-AHP (Analytic Hierarchy Process) combined weighting mechanism is introduced.

- The entropy weight method quantifies the "information contribution" of each feature.
- The AHP method incorporates domain experts' prior knowledge of "feature semantic value".
- This mechanism dynamically generates a feature weight vector, avoiding the limitations of traditional fixed weights or single subjective/objective weights.

(3)Experimental results on a Chinese cultural relic dataset demonstrate that the proposed method effectively captures complementary features and dynamically adapts to domain characteristics. Compared with baseline methods and improved approaches using "fixed-weight fusion," it achieves significant improvements in Precision, Recall, and F1-score.

Notably, the task addressed in this paper focuses on entity alignment within structured encyclopedic KGs—a distinction from traditional entity linking in unstructured text—thus providing a new perspective for heterogeneous data integration.

2. Method

This paper proposes a method of entity alignment based on multi-feature similarity fusion for Chinese entity alignment,

named MFSF-CEA. The entity attributes and abstract are important intrinsic characteristics of the entity, through which the entity characteristics at the level of words can be obtained. The entity-related context in encyclopaedias contains rich contextual information, which plays a certain role in determining the semantic distinction of entities. The fusion of entity context features obtains entity features from two levels of lexical and semantic relations. The entity alignment method based on multi-feature fusion is shown in Figure 1, including data preprocessing, feature obtaining and similarity calculation, multi-feature fusion, and entity alignment and output.

Firstly, the cultural relic data obtained from encyclopaedic websites are preprocesses, this involves deleting irrelevant information, stopping words, clauses, word segmentation, part-of-speech tagging, and syntactic analysis.

After data preprocessing, they are stored as a triple <entity attribute, entity abstract, entity context>, and entities in two encyclopaedic knowledge bases are formalized: $e_a =< attr, abst, docu >$ and $e_b =< attr, abst, docu >$, where e_a and e_b are the names of the entity related to cultural relics, $attr$ represents the entity attribute of cultural relics, $abst$ represents the entity abstract of cultural relics, and $docu$ represents the entity context information of cultural relics. For a cultural relic entity e_a, there are multiple attributes $attr_{ea}$, each of which has a corresponding attribute value $value_{ea}$. The entity abstract $abst$ is stored as text and forms the main content of entity description text. Entity context $docu$ is the unstructured text describing the entity on the web page, also known as entity context.

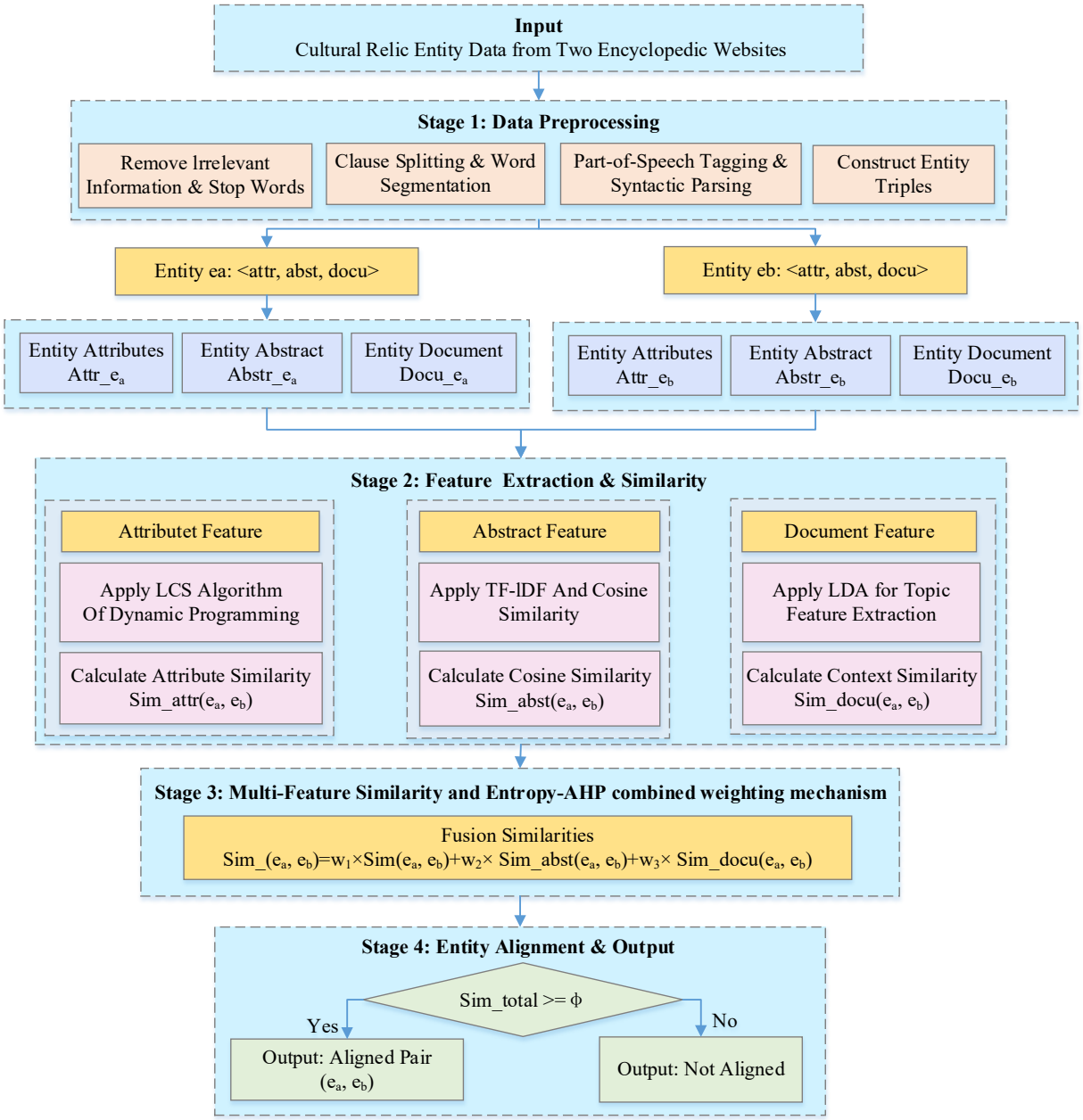

Figure 1. The framework structure of the MFSF-CEA model

Then, the features of entity attributes, entity abstract, and entity context are extracted, respectively, and the similarities are calculated. The LCS based on dynamic programming is used for entity attribute feature extraction and similarity calculation. The TF-IDF is used for entity abstract feature extraction and cosine similarity is used for similarity calculation. LDA is utilized to extract context information.

To address feature importance disparities in the cultural relic domain, we introduce an Entropy-AHP combined weighting mechanism. The entropy weight method objectively quantifies each feature's information contribution, while AHP incorporates domain experts' knowledge of feature semantic value. Then, the integrated similarity score $\text{sim}_{\text{final}}$ is compared against a threshold θ. Entity pairs exceeding this threshold are considered matches and added to the alignment set.

Finally, the set of entity alignments is the output.

2.1. Entity Attribute Similarity Calculation Based on the LCS of Dynamic Programming

The LCS algorithm is used to calculate entity attribute similarity. The LCS can distinguish whether two entities are similar at the character level. All the attribute values of an entity constitute a specific sequence of text. There are sequences of two entities, e_a and e_b. Without changing the relative positions of the elements, the result of removing zero or more elements from the sequence of two entities is called a subsequence of the sequence, and the longest common subsequence is a subsequence that is the longest of all the known common sub sequences.

Algorithm 1：LCS search algorithm based on dynamic programming	
Input:	The sequence of attribute values for two entities e_{ai} and e_{bj}
Output:	The LCS for e_{ai} and e_{bj}
Start:	
Step 1:	If $e_{ai} = e_{bj}$， because the LCS is discontinuous， the last element of the LCS for e_{ai} and e_{bj} is the current element and is the length of the LCS plus 1, i.e. $d[i][j] = d[i-1][j-1] + 1$.
Step 2:	If $e_{ai} \neq e_{bj}$， there are two cases for the value of d_{ij}.Take the highest value as the current value of A in both cases, i.e. the length of the LCS。 Case 1：The value of the $d[i-1][j]$ is the LSC length of the i-1 elements of e_{bi} and the j elements of e_{bj}. Case 2：The value of $d[i][j-1]$ is LSC length of the i elements of e_{ai} and the $j-1$ elements of e_{bj}.
Step 3:	Repeat Step 1 and Step 2 until take the last value of the matrix d.
End	

The attribute values of two entities e_a and e_b are expressed as $attr_{ea} = \{A_{a1}, A_{a2}, \dots, A_{am}\}$ and $attr_{eb} = \{A_{b1}, A_{b2}, \dots, A_{bn}\}$ respectively; the corresponding set of attribute values are $value_{ea} = \{v_{a1}, v_{a2}, \dots, v_{am}\}$ and $value_{eb} = \{v_{b1}, v_{b2}, \dots, v_{bn}\}$, where m and n are the number of attributes, A_{ai} represents the i_{th} attribute of the entity e_a, and A_{bj} represents the j_{th} attribute of the entity e_b. The number of characters of A_{ai} and A_{bj} is p and q, and the sequences of corresponding attribute values are, respectively, $v_{ai} = \{c_{a1}, c_{a2}, \dots, c_{ap}\}$ and $v_{bj} = \{c_{b1}, c_{b2}, \dots, c_{bq}\}$, where c_{ai} is the i_{th} character of the attribute value $v_{ai}(i > 0)$, c_{bj} is the j_{th} character of the attribute value $v_{bj}(j > 0)$. This paper attempts to calculate the similarity between attribute values of two entities by using the LCS algorithm to discriminate the similarity between two entities.

In the method for finding the LCS of two sequences, the simplest and most outrageous method is the violence enumeration method, which assumes that the sequence e_{ai} has p elements and the sequence e_{bj} has q elements, so e_{ai} and e_{bj} have 2^p and 2^q subsequences, respectively. If any two subsequences are compared one by one, the algorithm's complexity is 2^{p+q}. However, since the number of elements in the sequence is indefinite, i.e. the length of the sequences are indefinite, the algorithm's complexity increases exponentially with the sequence length, so this method is not suitable for long sequences [16].

2.1.1 Entity Attribute Similarity Calculation Based on the LCS of Dynamic Programming

To decrease the complexity of the algorithm, Knuth proposed a classical method based on dynamic programming for LCS selection, which decomposed the problem to be solved into several subproblems[17]. The subproblems were solved, the final solution was obtained from the solutions of the subproblems. This method reduced the complexity of the algorithm using the dynamic programming approach. Due to the uncertainty of the number of entity attributes in the encyclopaedic knowledge base, we use the LCS to calculate the similarity of entity attributes. At the same time, to cut down the complexity of the algorithm, this paper introduces a dynamic programming algorithm to extract the LCS.

The length of the entity attribute value sequence e_{ai} and e_{bj} are p and q; we then generate the matrix d with a size of $(p+1)\times(q+1)$ and the initial elements are all 0, where $d[i][j]$ represents the length of the LCS of the sequence e_{ai} and e_{bj}. The formula is as follows:

$$d_{ij}=\begin{cases}0 & if \quad i=0 \; or \; j=0\\ d[i-1][j-1]+1 & if \;\; i,j>0, e_{ai}=e_{bj}\\ max(d[i][j-1],d[i-1][j]) & if \;\; i,j>0, e_{ai}\neq e_{bj}\end{cases} \quad (1)$$

The algorithm 1 is finding the LCS of entity attributes based on dynamic programming. Although the traditional algorithm to obtain LCS using dynamic programming can transform the time complexity into $O(p*q)$, the storage of intermediate results adds space complexity into $O(p*q)$. Our method only uses the result of row $i-1$ and $d[i][j-1]$ before the next position. So, we compress the state into a one-dimensional array with space complexity $O(min(strlen(e_{ai}),strlen(e_{bj})))$, where the length of e_{ai} is p , and the length of e_{bi} is q.

Notably, our implementation optimizes space complexity to (min(p,q)) by only storing the current and previous rows of the DP table, which is particularly efficient for long sequences. This optimization significantly reduces memory consumption, making the LCS-based similarity calculation more scalable for large-scale entity alignment tasks.

2.1.2 Entity Attribute Similarity Calculation

After using the dynamic programming method to find the LCS of two entity attribute values, its similarity will be calculated to measure the similarity of entity attributes. The formula is as follows:

$$sim_{attr}(e_a,e_b)=\frac{LCS(e_{ai},e_{bj})}{max\,(len(e_{ai}),len(e_{bj}))}, \quad (2)$$

where $LCS(e_{ai},e_{bj})$ is the LCS of the sequence ea_i and eb_j of the attribute values of two entities. The similarity of entity attributes is better for distinguishing the similarity of two entities at the character level.

2.2. Entity Abstract Similarity Calculation Based on TF–IDF and Cosine Similarity

The abstract information of cultural relic entities in encyclopaedia knowledge bases contains a lot of potential corpus information, such as the abstract of "Afterglow-style 'Caifeng Mingqi' seven-stringed guqin" which includes the category of cultural relics, the museum's collection of information, and dynasty information, length and height, and the origin of its name of the "Afterglow-style 'Caifeng Mingqi' seven-stringed guqin". The similarity calculation of entity abstract information can have a certain influence on the similarity calculation results between two entities [18].

2.2.1 Feature Vector Acquisition Based on TF–IDF

Term Frequency–Inverse Document Frequency (TF–IDF) evaluates the importance of a word to a document in the corpus by counting the frequency of the word appearing in the text and the frequency of the word appearing in the corpus [19]. TF–IDF preserves the important words in the text and ignores the common but irrelevant one [20]. This paper uses TF–IDF to obtain the frequency of keywords in entity summary information, which obtains entity features from word level. Given that cultural relic descriptions are often short, we acknowledge potential limitations of TF-IDF in capturing semantic richness. To mitigate this, we incorporated n-gram features and refer to methods proposed in Chen et al. [21] for short text representation.

A vector space model is adopted in this paper to represent features of entity abstract, TF–IDF is used to obtain the weight of feature words, and Formula (3) is used to represent keyword feature vectors of the entity abstract. The keyword feature vector of entity abstract is shown as

$$V(d)=t_iw_i(d_j), \quad (3)$$

where d_j represents the text in the corpus and t_i is the feature word of text $d_j, i\in\{1,2,\dots,n\}$. Text d_j contains n feature words t and n corresponding weights w . w_i represents the weights of the feature word t_i in the text d_j, which is the frequency of t_i appearing in text d_j, and is calculated as follows:

$$tf_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}, \quad (4)$$

$$idf_i=log\frac{|D|}{1+|D_{t_i}|}, \quad (5)$$

$$tfidf_{i,j}=tf_{i,j}\times idf_i, \quad (6)$$

where $tf_{i,j}$ represents the term frequency of the feature word t_i, $n_{i,j}$ is the number of times t_i appears in text d_j, $n_{k,j}$ is the sum of all words appearing in text d_j, idf_i represents the inverse document frequency, $|D|$ is the number of texts in the corpus, and $|D_{ti}|$ is the total number of words t_i in text d_j. To avoid words that do not appear in the corpus, that is, when the total number of words is 0, this paper uses $1+|D_{ti}|$ as the denominator. $tfidf_{i,j}$ represents term frequency–inverse document frequency, which is the product of the term frequency $tf_{i,j}$ and the inverse document frequency idf_i. The greater the value of $tfidf_{i,j}$, the more important the feature t_i is in the corpus text d_j.

2.2.2 Entity Abstract Similarity Calculation

The feature vector of the entity abstract is obtained by TF–IDF algorithm after obtaining the weight of the feature word. In this paper, the cosine vector similarity measurement method is used to measure the similarity of the abstract between two entities. Cosine similarity is a word-based similarity calculation method to determine whether two entities are similar at the level of words. For the n dimensional

vectors of entities e_a and e_b, the cosine similarity between them can be calculated as follows:

$$Sim_{abst}(e_a, e_b) = \frac{\sum_{i=1}^{n} e_{a_i} \times e_{b_j}}{\sqrt{\left(\sum_{i=1}^{n}\left(e_{a_i}\right)^2\right)} \times \sqrt{\left(\sum_{i=1}^{n}\left(e_{b_j}\right)^2\right)}} \quad (7)$$

2.3 Entity Context Similarity Calculation Based on LDA

In encyclopaedic websites, unstructured data usually contains multidimensional description information of entities, and most of the description information is presented in text form. Latent Dirichlet Allocation (LDA) is a particularly popular method for fitting a topic model, which is robust for large text sets [22]. For the cultural relic entities of encyclopaedia websites, this paper uses LDA to conduct topic models for context information and extract topic features, to obtain the similarity features of the entities from the context semantic level and determine whether the entities are aligned through the extracted topic feature distribution.

The similarity calculation process of entity context based on the LDA topic model includes three steps, topic modelling based on LDA model, topic feature generation based on entity context, and entity context similarity calculation.

2.3.1 Topic Modelling Based on the LDA Model

LDA extracts document topics from the document collection, calculates the probability distribution of each document topic, and conducts topic clustering according to the topic's distribution. LDA for topic modelling considers each document to be a set of topics arranged in a certain proportion, and each topic is a set of keywords in a certain proportion. Each document can contain a specific proportion of words for the topic. Words make up topics, and topics make up documents. The document-to-topic distribution follows Dirichlet, and topics-to-words follow a polynomial distribution. Every word is produced by a topic; each topic is the probability distribution of all the words. Words with the same topic pair are given a higher probability. LDA can rearrange the distribution of topics within the documents and keywords when the number of topics is provided.

The LDA topic model is a three-tier Bayesian probability model in Figure 2.

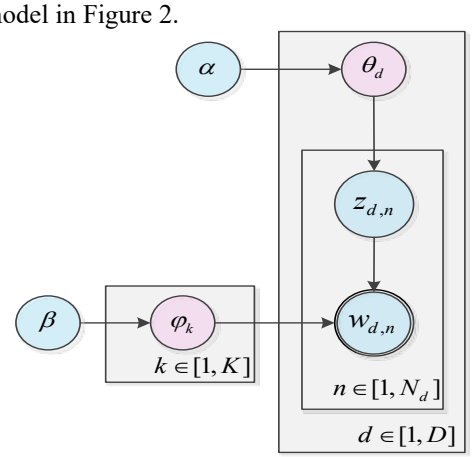

Figure 2. The process of the LDA model generates entity context

The basic idea of LDA is to calculate text similarity by topic modelling, the words in the text are extracted by traversing the word distribution corresponding to the topic, and text similarity is calculated by the word distribution. A topic is generated according to the topic probability θ_d of a particular document, which in turn, generates each observed word at the index of the document d based on the polynomial z_{ij} of the words of a particular topic. Two polynomials z_{ij} and $\varphi_{z_{ij}}$ are produced by two Dirichlet distributions with hyperparameters α and β, respectively. For the entity e of a cultural relic in the encyclopedia knowledge base, the detailed process of generating a document by the LDA model is described as follows:

(1) Sampling from the Dirichlet distribution α generates a topic polynomial distribution matrix θ_d for entity context documents, where each column represents the probability of each topic appearing in each document.

(2) Sampling from the topic polynomial distribution θ_d generates the topic z_{ij} of the j_th word in the entity context document i.

(3) Sampling the Dirichlet distribution β generates the word distribution $\varphi_{z_{ij}}$ corresponding to the entity context topic z_{ij}.

(4) Sampling from the word polynomial distribution $\varphi_{z_{ij}}$ generates the feature word w_{ij}.

2.3.2 Entity context topic feature generation

In the LDA model, the topic of the text is represented as an implicit variable, that is, both θ and φ are unknown, which can be derived through the probabilistic derivation method of Gibbs sampling [23-24]. In this paper, Gibbs sampling is used to estimate the unknown parameters and solve the topic model.

(1) Document–topic probability matrix

After obtaining the entity context information from the encyclopaedia page, the document–topic probability matrix is obtained through topic modelling of the contextual information of the entity and the estimation of the implicit variables in the model,

$$\theta_d = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1k} \\ P_{21} & P_{22} & \dots & P_{2k} \\ & & \dots & \\ P_{d1} & P_{d2} & \dots & P_{dk} \end{bmatrix}_{d \times k} \quad (8)$$

where P_{dk} represents the probability that the k_th topic will be included in the d_th document, d is the number of documents in the dataset of the entity document, and k is the number of topics in the document. The probability values $(P_{i1}, P_{i2}, \dots, P_{in})$ for each row in the matrix θ_d are sorted in reverse order to get the values $(P'_{i1}, P'_{i2}, \dots, P'_{in})$, and the inverse matrix θ'_d is obtained by sorting the probability values of each row in the matrix θ_d. N topics with the highest probability of generating topics in each document are

obtained through the matrix after inversion. The feature word set and feature matrix are obtained by taking $n(n < c)$ words in k topics as follows, $n(n < k)$ topics in d documents are taken to obtain the topic feature set and feature matrix.

$$z_{dn} = \{k'_0, k'_1, \cdots, k'_n\} \quad (9)$$

$$feature(k_{dn}) = \begin{bmatrix} P'_{11}\ P'_{12} \dots P'_{1n} \\ P'_{21}\ P'_{22} \dots P'_{2n} \\ \dots \\ P'_{d1}\ P'_{d2} \dots P'_{dn} \end{bmatrix}_{d \times n} \quad (10)$$

The maximum value is found in the feature matrix $feature(k_{dn})$ as the feature value V'_{ki} of each topic k'_i in the topic feature set z'_{dn}, and the topic feature vector can be generated as follows:

$$V_{ek} = \{v'_{k0}, v'_{k1}, \cdots, v'_{kn}\} \quad (11)$$

(2) Topic–word probability matrix

By modelling the entity context and using Gibbs sampling for probabilistic derivation solution, the topic–word probability matrix is obtained as follows,

$$\varphi_{z_{kc}} = \begin{bmatrix} P_{11}\ P_{12} \dots P_{1c} \\ P_{21}\ P_{22} \dots P_{2c} \\ \dots \\ P_{k1}\ P_{k2} \dots P_{kc} \end{bmatrix}_{k \times c} \quad (12)$$

where P_{kc} is the probability that the c_th word belongs to the topic k, and c is the number of total words. The probability values $(P_{i1}, P_{i2}, \dots, P_{ic})$ for each row in the matrix $\varphi_{z_{kc}}$ are sorted in reverse order to get the values $(P'_{i1}, P'_{i2}, \dots, P'_{ic})$, and the inverse matrix $\varphi'_{z_{kc}}$ is obtained by sorting the probability values of each row in the matrix $\varphi_{z_{kc}}$. C words with the highest generation probability of each topic are obtained through the inverse matrix. The feature word set and feature matrix are obtained by taking $n(n < c)$ words in k topics as follows:

$$W_{kn} = \{w'_0, w'_1, \cdots, w'_n\}, \quad (13)$$

$$feature(w_{dn}) = \begin{bmatrix} P'_{11}\ P'_{12} \dots P'_{1n} \\ P'_{21}\ P'_{22} \dots P'_{2n} \\ \dots \\ P'_{k1}\ P'_{k2} \dots P'_{kn} \end{bmatrix}_{k \times n} \quad (14)$$

where n represents unrepeated feature words. For the word w'_i in the feature word set W_{dn}, its feature value is the maximum value in its feature matrix $feature(w_{kn})$, and the word feature vector is generated as follows:

$$V_{ec} = \{v'_{c0}, v'_{c1}, \cdots, v'_{cn}\} \quad (15)$$

2.3.3 Entity Context Similarity Calculation

The context similarity of each entity is calculated by using the cosine similarity through the topic feature vector and feature word vector of each entity. The similarity calculation method is shown as follows:

$$sim(ek_i, ek_j) = \frac{v_{ki} \cdot v_{kj}}{|v_{ki}|\,|v_{kj}|} \quad (16)$$

$$sim(ec_i, ec_j) = \frac{v_{ci} \cdot v_{cj}}{|v_{ci}|\,|v_{cj}|} \quad (17)$$

$$sim_{docu}(e_a, e_b) = sim(ek_i, ek_j) + sim(ec_i, ec_j) \quad (18)$$

where $sim(ek_i, ek_j)$ is the topic feature similarity of the document, $sim(ec_i, ec_j)$ is the similarity of the feature words for the topic, and $sim_{docu}(e_a, e_b)$ represents the entity context similarity.

2.4 Model and Algorithm of Entity Alignment

2.4.1 Entity Alignment Based on Multi-Feature Similarity and Entropy-AHP combined weighting mechanism

To address the limitations of fixed-weight fusion strategies, which fail to account for the varying importance of different features in the cultural relic domain, this paper introduces an Entropy-AHP combined weighting mechanism. This mechanism dynamically determines the optimal weights for integrating multi-granularity features. The entity alignment model is formally defined as follows:

$$Sim(e_a, e_b) = w_1 \times Sim_{attr}(e_a, e_b) + w_2 \times Sim_{abst}(e_a, e_b) + w_3 \times Sim_{docu}(e_a, e_b). \quad (19)$$

In Formula 19, $Sim(e_a, e_b)$ represents the similarity calculation model of two entities after multi-feature fusion, which is calculated by the weighted sum of the similarity of multiple features. $Sim_{attr}(e_a, e_b)$ is entity attribute similarity, $Sim_{abst}(e_a, e_b)$ is entity abstract similarity, $Sim_{docu}(e_a, e_b)$ is entity context similarity; w_1, w_2 and w_3 are weights of entity attribute similarity, entity abstract similarity, and entity context similarity, respectively. The value range of parameters $w_1, w_2, w_3 \in [0,1]$ and $w_1 + w_2 + w_3 = 1$.

The Entropy-AHP combined weighting mechanism operates in two stages:

(1) Objective Quantification (Entropy Weight Method): This stage calculates the information entropy of each feature to assess its discriminative power within the dataset. A feature with lower entropy (e.g., "cultural relic age") indicates higher discriminability and is assigned a higher weight autonomously.

(2) Subjective Correction (Analytic Hierarchy Process): This stage incorporates domain expertise to evaluate the

inherent "semantic value" of features. For instance, attributes like "excavation site" are prioritized over "restoration history" based on their relative importance for identifying cultural relics, as determined by expert judgment.

The final weight for each feature is obtained by combining the objective entropy weight and the subjective AHP weight, ensuring the fusion process is both data-driven and semantically aligned with domain-specific knowledge.

2.4.2 Entity Alignment Algorithm Based on Multi-Feature Similarity and Dynamic Weighting

Building upon the proposed model, we design an entity alignment algorithm that integrates entity attributes, abstracts, and context information. The core of this algorithm lies in its dynamic feature fusion strategy, which calculates entity similarities from multiple perspectives and achieves alignment by comparing the integrated similarity score against a threshold φ. The input to the algorithm is a corpus from Baidu Baike and Chinese Wikipedia, structured as triples <entity attribute, entity abstract, entity context>. The detailed procedure of the entity alignment algorithm is outlined as follows.

Algorithm 2：Entity alignment algorithm based on multi-features similarity and Dynamic Weighting	
Input:	Entity corpus of cultural relics from encyclopedia knowledge base include entity attribute, entity abstract and entity context.
Output:	Alignment entity set
Start:	
Step 1:	Obtain the textual content of cultural relics entity from the encyclopedia site, and store in the form of a triple < entity attribute, entity abstract, entity context >;
Step 2:	Data pre-processing. The tags and special symbols in the acquired corpus are filtered. Perform word segmentation and remove stop words, etc.;
Step 3:	Compute the entity attribute similarity $Sim_{attr}(e_a, e_b)$ using LCS of dynamic programming;
Step 4:	Compute the entity abstract similarity $Sim_{abst}(e_a, e_b)$ using TF-IDF and cosine similarity;
Step 5:	Compute the entity context similarity $Sim_{docu}(e_a, e_b)$ using LDA topic model;
Step 6:	Based on the Entropy-AHP combined weighting mechanism, calculate the dynamic weights w_1, w_2, w_3 for the three features: entity attributes, abstracts, and context;
Step 7:	Construct the multi-feature model of entity alignment algorithm, which combinates the weighted sum of the entity attribute similarity $Sim_{attr}(e_a, e_b)$, the entity abstract similarity $Sim_{abst}(e_a, e_b)$ and Calculate the entity context similarity $Sim_{docu}(e_a, e_b)$. Then obtain the Multi - feature similarity $Sim(e_a, e_b)$.
Step 8:	Set a threshold φ， When the similarity of two entities $Sim(e_a, e_b) > \varphi$， two entities e_a and e_b can be aligned, otherwise the two entities will not align.
Step 9:	Output an aligned collection of entities.
End.	

3. Experiments

3.1 Datasets

To verify the validity of the proposed alignment method, the dataset required for the experiment is constructed.

In this paper, five categories of cultural relics such as pottery, bronze, gold, silver, jade and porcelain are selected, in which we choose three categories of entities such as cultural relics' name (CRN), unearthed location (UL), and museum collection (MC) to form the dataset. Due to the heavy workload and low accuracy of manual annotation data, this paper extracts entities from the Baidu Encyclopaedia and Chinese Wikipedia to automatically generate training data sets by using heuristic rules with the help of some structured information in the encyclopaedia.

The heuristic rules for positive example selection in the training dataset are as follows: (1) Two article entries have the same and unique names, that is, the two article entries have no synonyms or aliases. (2) Two articles with the same title, and the content similarity of the article must exceed 95%. (3) The category labels of the two articles are exactly the same, and the content similarity of the two articles is over 95%. (4) Two articles have the same title and the same category tag, they will be listed as referring to the same entity.

In addition, there is the function of synonym digestion in the encyclopaedia. That is, an article name has an alias but will be automatically replaced with the standard entity name when looked up on the encyclopaedia site, e.g., "sham Li Bo"

is the alias of “Shaanxi history museum”. When "sham Li Bo" is received as input, it will be replaced by the “Shaanxi History Museum” automatically, and the information about the Shaanxi Provincial History Museum is returned.

Many rules are used for negative example selection in the training dataset. Except for rules that generate positive examples, all other rules can generate negative examples. Enumerate the main heuristic rules used for negative example generation. (1) The titles of two articles are different, and the content similarity of the two articles is less than 50%. (2) The category labels of the two articles are completely different, and the content similarity of the two articles is less than 50%. (3) The two articles are in different fields.

The descriptions of three types of cultural relic entities, cultural relic’s name (CRN), unearthed location (UL), museum collection (MC), the experimental dataset are shown in Table 1. This paper divides the dataset into two disjoint parts. A development set to obtain the values of the parameters of the algorithm. A test set to measure the performance of the algorithm. In the dataset of cultural relic entities, the development set accounts for 40%, and the test set accounts for 60%.

Table 1. The dataset of cultural relic entities

Types	Baidu Encyclopaedia	Wikipedia	positive example	negative example
CRN	2000	2000	5284	2851
MC	1600	1600	2846	1062
UL	1600	1600	4972	2038

3.2 Experiment Settings

Evaluation. In this paper, classification precision (P), recall (R), and F1-Score are used as model performance evaluation indicators.

Baselines. To prove the effectiveness of MFSF-CEA proposed in this paper, the experimental results are compared with the following baselines and two groups of comparative experiments were set up. The first group is compared with five typical algorithms: Chen et al. proposed an integrated framework for multi-source heterogeneous data fusion intended for the reconstruction and expansion of long-line expressways [8]. Wang et al. proposed a multi-source data fusion method based on the expanded vector space model. [10]. Akhtar et al. proposed a multi-source data fusion method based on the expanded vector space model [13]. Singh et al. proposed a similarity-based semi-supervised algorithm for automatically labelling unlabelled text data [18]. Zhao et al. achieved multi-modal knowledge graph completion by aligning multi-level semantics [19]. The second group is the model that fuses different feature similarity, that is, based on the entity attribute, entity abstract, and entity context similarity calculation proposed in this paper, different features or feature combinations are adopted to conduct experimental comparisons with the method proposed in this paper.

(3) Experimental setup. This paper set $\alpha = 0.1$ and $\beta = 0.1$. In the entity alignment model with multi-feature similarity fusion, the weight values of w_1, w_2 and w_3 are determined by the importance degree of each feature in the model and compared through multiple experiments. This paper set $w_1 = 0.4$, $w_2 = 0.2$ and $w_3 = 0.4$. The entity alignment algorithm threshold φ is directly related to the effect of the setting and the real alignment. The higher the threshold value φ, the greater the accuracy of entity alignment. However, when the value of threshold φ is set too low, the corresponding recall rate decreases and recall increases, but accuracy goes down. In this paper, the optimal value of F1-score in the experiment is used to set the threshold value φ, and the threshold φ with high accuracy is taken when F1-scores are similar.

Furthermore, the threshold φ in the entity alignment algorithm was set to a unified value, which was determined by optimizing the F1-score on the development set.

This study employs a unified similarity threshold φ across all cultural relic entity categories to simplify configuration and ensure operational feasibility. We recognize that different entity types may, in theory, have different optimal decision boundaries due to variations in their descriptive text feature distributions. However, in the absence of prior knowledge about entity types and while pursuing a general solution, adopting a unified threshold that performs optimally on the overall development set represents a reasonable and robust choice, guaranteeing baseline performance for unseen entity types.

3.3 Experimental Results and Analysis

3.3.1 Comparison Experiment of Entity Alignment for Cultural Relics

To verify the overall performance of MFSF-CEA proposed in this paper in the task of entity alignment for cultural relics, the comparative experiments between the MFSF-CEA model and the first group of classical algorithms are executed using test set. In this paper, three main entity types (CRN, UL, MC) are selected for the comparative experiments. The experimental results are shown in Table 2

Table 2: Experimental results of entity alignment for cultural relics

Methods	CRN (%)	MC (%)	UL (%)

	P	R	F1	P	R	F1	P	R	F1
Chen et al. [8]	66.18	62.96	64.53	71.31	68.28	69.76	68.49	65.38	66.90
Wang et al.[10]	79.25	65.49	71.72	75.38	73.02	74.18	76.02	67.17	71.32
Akhtar et al. [13]	79.72	68.34	73.59	81.04	76.82	78.87	78.69	71.56	74.96
Singh et al.[18]	82.34	75.00	78.50	83.23	77.73	80.38	83.81	76.83	80.17
Zhao et al. [19]	84.51	77.00	80.58	86.12	84.89	85.50	83.27	77.44	80.25
MFSF-CEA	90.62	87.11	88.83	93.23	92.13	92.67	92.24	88.57	90.37

As shown in Table 2, Chen et al.'s algorithm performed poorly in the entity alignment task for cultural relics by relying solely on the feature template of cultural relic attributes. The F1 values of the three types of entities were 64.53%, 69.76%, and 66.90%, respectively. MFSF-CEA particularly excels when entities share rich contextual and structural features (e.g., tomb artifacts with detailed excavation records), whereas Zhao et al.'s method struggles with sparse attribute overlap. This was mainly due to the particularity of word-formation of cultural relic entities, and the fact that the feature template could not cover all the features of the entity attributes comprehensively. Zhang et al.'s algorithm performed relatively well by using the LDA theme model to weigh multiple features of entities, which reflected the effectiveness of the LDA theme model in entity feature extraction. For the three types of entities (CRN, UL, MC), the MFSF-CEA method is higher than the comparison model in terms of accuracy, recall, and F1-score. From the comparison results, we can see that the MFSF-CEA method is effective in the entity alignment task for cultural relics in the encyclopaedic knowledge base.

In addition, high-accuracy alignment results should be obtained in the entity alignment task for cultural relics, so that the aligned entities can be directly integrated with the knowledge base to build a high-quality knowledge graph. Therefore, the classification threshold is set relatively high in the experiment, resulting in the accuracy of MFSF-CEA being relatively higher than the recall. Due to the mutual restriction between precision and recall, this paper further measured the experimental results by calculating the average F1-score. Overall, the F1-scores of MFSF-CEA for the three types of entity in this paper are 78.98%, 85.60%, and 82.15% respectively, and the method achieves optimal performance. This is mainly because this paper integrates multiple features of entities and applies effective feature extraction methods to realize entity alignment according to the characteristics of different features of entities.

3.3.2 The Influence of the Number of Topics on the Experimental Results

When the LDA topic model captures the topic and word characteristics of entity context information, the number of topics K needs to be set. In this paper, a series of experiments are conducted to set different numbers of topics for the three types of entity using development set. The experimental results are shown in Figure 3. The subgraphs (a), (b), and (c) represent the entity alignment results of CRN, UL, and MC under different numbers of topics, respectively.

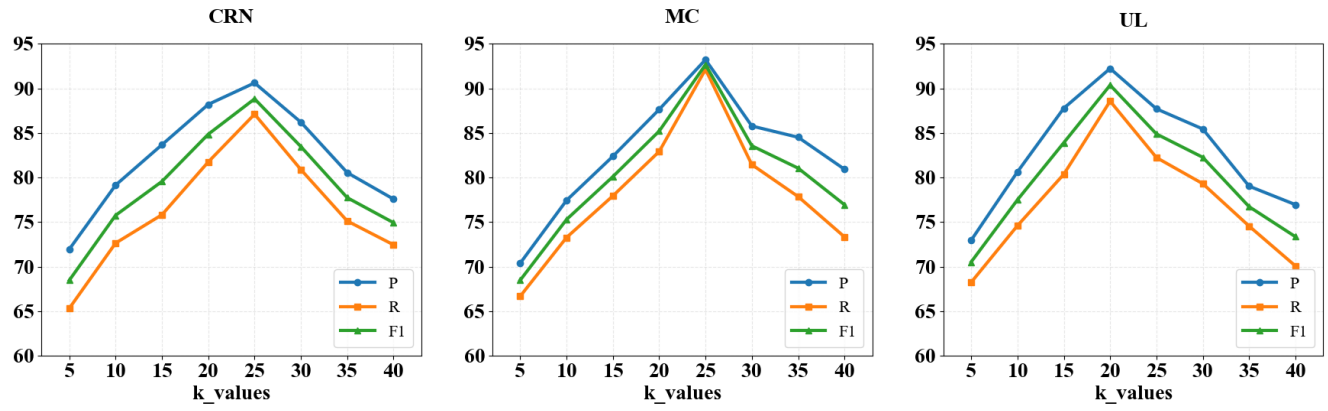

Figure 3. The influence of the number of topics on the experimental results

As can be seen from the experimental results in Figure 4, when the number of topics K is between 4 and 6, especially when K=5, the result of each evaluation indicator is the highest. Therefore, in the experiment, the number of topics K=5 is selected in this paper. Relatively, the alignment results of MC are higher than the alignment results of CRN and UL. This is mainly because the MC description information is relatively professional and most of them are proper nouns, the number of entities to be aligned is relatively small, and the alignment effect is

relatively strong. In addition, the formation characteristics of CRN and UL, and the fact that the knowledge descriptions in the encyclopaedic knowledge base are not accurate enough, affect the model's performance when obtaining the characteristics of these two types of entities. This is also the research direction that needs to be optimized in the later stage of this paper.

3.3.3 Influence of Entity Features on the Entity Alignment Performance for Cultural Relics

To verify the effect of the multi-feature fusion method on the entity alignment performance for cultural relics, we set the comparison experiment between the MFSF-CEA method and the second group with different feature fusion using test set. Entity attribute, entity abstract, and entity context are taken as three single features, respectively, and two features are combined to realize the entity alignment of cultural relics. In this paper, three main entity types are selected for comparative experiments. The experimental results are shown in Table 3.

Table 3. The influence of entity features on the entity alignment performance for cultural relics

Methods	CRN (%)			MC (%)			UL (%)		
	P	R	F1	P	R	F1	P	R	F1
Attribute	73.56	70.71	72.11	75.68	74.79	75.23	74.88	71.90	73.36
Abstract	78.63	75.58	77.08	80.89	81.02	80.95	80.04	76.85	78.41
Context	76.85	73.88	75.33	79.07	78.13	78.60	78.23	75.11	76.64
Attribute_Abstract	82.65	79.45	81.02	85.04	84.03	84.53	84.13	80.78	82.42
Attribute_Context	87.37	83.98	85.64	89.88	88.82	89.35	88.93	85.39	87.12
Abstract_Context	85.39	82.08	83.70	87.85	86.81	87.33	86.92	83.46	85.15
MFSF-CEA	90.62	87.11	88.83	93.23	92.13	92.67	92.24	88.57	90.37

The experimental results are shown in Table 3. The results indicate that methods relying on a single feature generally underperform. Notably, for the UL (Unearthed Location) entity type, the combination of Attribute and Abstract features yielded a lower F1-score than using Context features alone (see Table 3). This can be attributed to the particularity of UL entity descriptions: the semantic context often contains richer and more discriminative information about excavation sites and historical background, which is better captured by the LDA topic model than by the combination of structured attributes and sparse abstract keywords.

It is worth noting that the multi-feature fusion method (MFSF-CEA) based on entity attributes, entity abstract, and entity context achieves the optimal effect, which reflects the effectiveness of comprehensively integrating multi-granularity features for the entity alignment task of cultural relics.

3.3.4 Effectiveness of the Entropy-AHP Dynamic Weighting Mechanism in Entity Alignment

To systematically evaluate the effectiveness and rationality of the Entropy-AHP dynamic weighting mechanism, this experiment compares it against four weighting strategies: Average Weighting (baseline), Empirical Weighting (fixed optimization), Entropy Weighting only (objective data-driven), and AHP Weighting only (subjective knowledge-driven). All methods utilize the same underlying features (LCS, TF-IDF, LDA), differing only in the weight generation approach, to fairly validate the proposed fusion mechanism's role in enhancing entity alignment performance and weight allocation rationality.

(1) Overall Performance Comparison

All comparative methods were executed on a unified test set, and their performance metrics were recorded in table 4.

Table 4. Impact of different weighting strategies on cultural relic entity alignment performance

Methods	CRN (%)			MC (%)			UL (%)		
	P	R	F1	P	R	F1	P	R	F1
Average Weighting	67.79	65.16	66.45	69.74	68.92	69.33	69.00	66.25	67.60
Empirical Weighting	72.89	70.07	71.45	74.99	74.11	74.55	74.19	71.24	72.69
Entropy Weighting	78.38	75.34	76.83	80.63	79.68	80.16	79.78	76.60	78.16
AHP Weighting	84.28	81.01	82.61	86.70	85.68	86.19	85.78	82.37	84.04

MFSF-CEA (Entropy-AHP)	90.62	87.11	88.83	93.23	92.13	92.67	92.24	88.57	90.37

As shown in the table 4, the proposed Entropy-AHP method significantly outperforms all baseline models across all metrics. This demonstrates that the dynamic weighting strategy, which combines objective statistics with domain prior knowledge, can guide multi-feature fusion more effectively than any single strategy alone.

(2) Weight Analysis and Interpretability Validation

To gain deeper insight into why Entropy-AHP is effective, we extract and analyse the final weight vector it generates. The experimental results are shown in Figure 4.

High-Weight Features: Attribute similarity received the highest weight (0.48), which aligns with common knowledge in the cultural relic domain—structured attributes such as the name, era, and excavation site of a relic are highly discriminative features for distinguishing different entities.

Low-Weight Features: Abstract similarity received the lowest weight (0.24), consistent with our earlier findings in the "Discussion" section that keywords in abstracts suffer from sparsity issues and have weaker discriminative power. The Entropy-AHP mechanism successfully automatically suppressed the contribution of such features.

Mechanism Synergy: The final weights incorporate both the objective patterns learned from the data by the entropy weight method (e.g., high discriminability of attribute features) and the prior knowledge from domain experts incorporated by the AHP method (e.g., the importance of contextual topic semantics). This achieves a unification of data and knowledge, reflecting the method's rationality and interpretability.

Through the systematic experiments above, we can draw the following conclusions:

- Performance Improvement: The Entropy-AHP combined weighting mechanism significantly improves the precision, recall, and F1-score of entity alignment through dynamic and adaptive fusion of multiple features.
- Rational Decision-Making: The mechanism can automatically identify and enhance the contribution of key features (e.g., cultural relic attributes) while suppressing the influence of features with weaker discriminative power. Its decision-making process highly aligns with domain logic, demonstrating good interpretability.

This experiment fully proves the value of introducing a dynamic weighting mechanism into the multi-feature entity alignment framework, providing an effective solution to the problem of uneven feature importance in vertical domains.

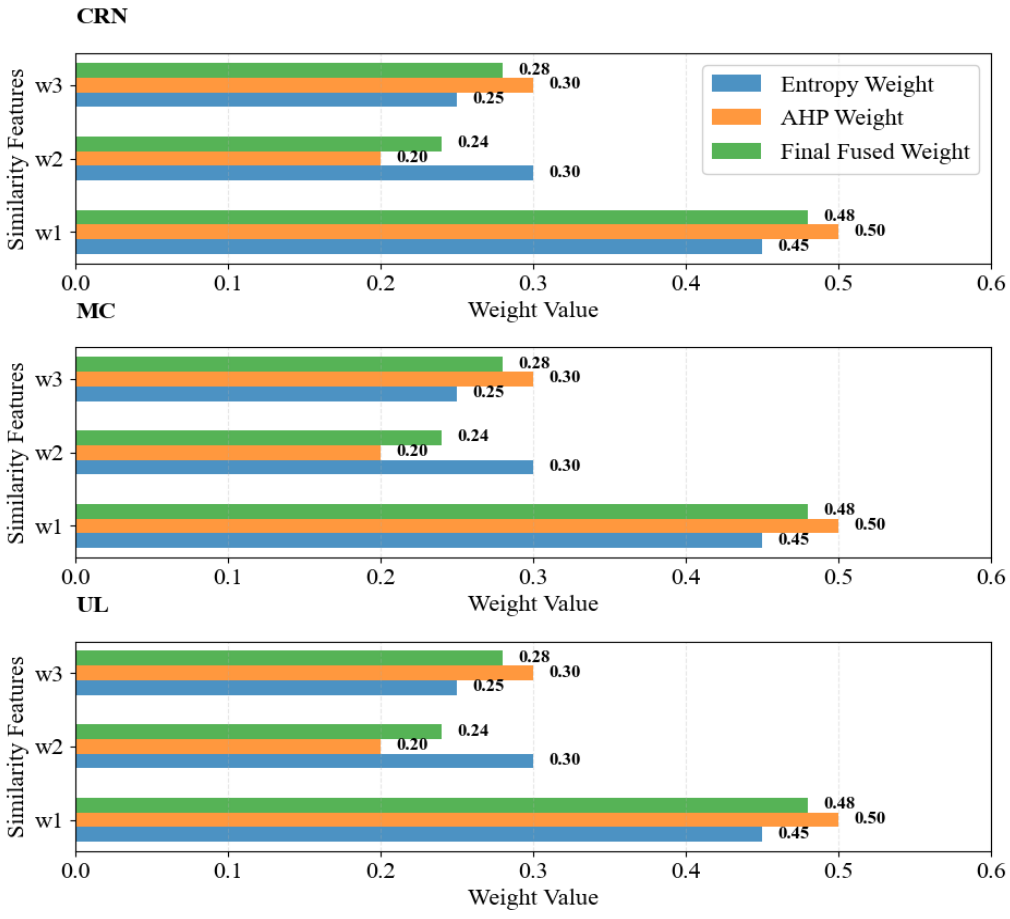

Figure 4. Weight Analysis and Interpretability Validation

4. Discussion

The findings of this study validate the design rationale of the MFSF-CEA model and reveal the distinctive characteristics of entity alignment in vertical domains. More importantly, by comparing our results with existing research, our findings both corroborate certain prior conclusions and offer new insights, thereby highlighting the advantages of our proposed method.

First, this study found that entity alignment relying solely on word-level abstract features (TF-IDF) yielded the lowest performance. This finding echoes the conclusions of Wang et al. [10], who also noted that the performance of keyword-based models significantly decreases when entity descriptions are sparse. However, our research extends this understanding further: in the specific vertical domain of cultural heritage, the descriptive texts of cultural relic entities often contain a multitude of non-standardized terms and classical language, making traditional text features like TF-IDF even more challenging. Thus, our study not only confirms the existence of this issue in the cultural heritage domain but also explains the domain-specific reasons behind it.

Second, the strong performance of sentence-level contextual features (LDA) as a standalone component strongly supports the proposition by Zhao et al. [19] regarding the use of deep semantic relations to enhance alignment performance. While their work emphasized the importance of cross-modal semantic alignment, our experiments demonstrate that topic-level semantic coherence, captured via LDA, is also crucial for accurate entity alignment even within purely textual cultural relic knowledge graphs. This complements existing research by indicating that topic models serve as an effective semantic enhancement tool in vertical domains rich in unstructured textual information.

However, our findings also diverge from some previous studies. For instance, the method proposed by Chen et al. [8], which relied heavily on manually constructed attribute feature templates, performed well in their specific experimental setting. Yet, our experiments show its performance is suboptimal in large-scale and highly heterogeneous encyclopaedic knowledge bases. This highlights the limitations of relying solely on predefined rules or templates in complex real-world scenarios, thereby underscoring the necessity of our proposed adaptive weighting mechanism.

Most critically, the overall performance of our proposed multi-granularity feature fusion framework based on Entropy-AHP significantly surpasses that of state-of-the-art models like Akhtar et al. [13] and Singh et al. [18]. Akhtar et al.'s model mines relational semantics through complex neural networks, while Singh et al. focus on similarity computation under a semi-supervised paradigm. Our approach does not seek to replace these deep semantic mining techniques but rather provides a complementary solution: an interpretable, dynamic weighting mechanism that integrates both objective data and subjective domain knowledge to optimize the synergy between features of different granularities. Experiments confirm that this strategy, which combines character-level exact matching (LCS), word-level distribution (TF-IDF), and sentence-level topic semantics (LDA, can more comprehensively address the complexity inherent in cultural relic entity descriptions.

In summary, through systematic comparison with prior work, this study not only validates the effectiveness of multi-granularity feature fusion for the entity alignment task in cultural heritage but, more importantly, elucidates how the Entropy-AHP dynamic weighting mechanism addresses the universal challenge of uneven feature importance in vertical domains by harmonizing data-driven patterns with domain prior knowledge, offering an effective and interpretable solution.

5. Conclusions and Future Work

This paper has introduced MFSF-CEA, a novel entity alignment method grounded in multi-feature similarity fusion, to address the critical challenge of integrating entities from heterogeneous sources in vertical domain knowledge graphs. The core of our approach lies in its dual-layer optimization framework encompassing multi-granularity semantic modelling and domain-adaptive dynamic fusion, which systematically captures and integrates complementary features from character, word, and sentence levels.

Experimental results on the domain-specific task of Chinese cultural relic entity alignment validate the effectiveness of our model. The integration of character-level attribute similarity (via LCS), word-level abstract similarity (via TF-IDF), and sentence-level contextual semantic similarity (via LDA) creates a synergistic effect, enabling accurate identification of equivalent entities despite variances in surface form and descriptive focus. More importantly, the proposed Entropy-AHP combined weighting mechanism demonstrates significant advantages over static fusion strategies by dynamically adapting to domain-specific characteristics.

The contributions of this work extend beyond technical performance. Methodologically, we establish a holistic and interpretable fusion framework that advances entity alignment from traditional string matching toward deeper semantic integration, providing an extensible solution paradigm for handling feature importance disparities in vertical domains. Practically, this research offers concrete technological support for digital humanities and cultural

heritage preservation, laying a solid foundation for constructing large-scale, high-precision knowledge graphs of Chinese cultural relics.

Although the MFSF-CEA model has demonstrated excellent performance in the task of entity alignment for Chinese cultural relics, this study has certain limitations that point to clear directions for future research. A core innovation of our model is its dynamic weighting mechanism, specifically designed for the characteristics of the Chinese cultural relics domain. However, this high degree of domain adaptation may, to some extent, constrain its cross-domain generalizability. The model's performance relies heavily on the specific style of the domain text and the availability of expert knowledge. The current model focuses on mining textual semantic features (character, word, text) from entity descriptions. However, the rich structural information within knowledge graphs (such as attribute graph relationships and category hierarchies between entities) has not yet been incorporated into the fusion framework.

Based on the aforementioned limitations, we plan the following future research directions to further enhance the model's generalizability, depth of information utilization, and practical applicability. Enhancing Cross-Domain Generalizability: We will explore meta-learning or domain adaptation techniques. The goal is to learn a universal feature weight prior from multiple vertical domains (such as traditional Chinese medicine or classical Chinese poetry), enabling the model to quickly adapt to the data distribution of new domains and reduce its reliance on domain expert knowledge and specific textual styles. Integrating Multimodal and Structural Information: Future work will move beyond purely textual semantics by striving to build a multimodal fusion framework. We plan to integrate the graph structural information from the knowledge graph (e.g., modelling attribute relationships between entities using Graph Neural Networks) and even the visual image features of cultural relics into the alignment model, thereby constructing a more comprehensive and robust evidence system for alignment.

Acknowledgements

This research was funded by the Ministry of Education in China Project of Humanities and Social Sciences, grant number 23YJA870016, Project Title: Research on Intelligent Question Answering and Knowledge Recommendation Systems for Smart Museums Based on Cultural Relic Knowledge Graphs; General Project of Specialized Research on Philosophy and Social Sciences in Shaanxi Province, Project Title: Research on the Empowerment of Intelligent Navigation and Cultural Dissemination Paths of Shaanxi Museum by Cultural Relics Knowledge Graph. The National Social Science Fund of China, grant number 23BTY025. The Key (Supported) Discipline of Shaanxi Xueqian Normal University.

References

[1] Simos MA, Makris C. Computationally efficient context-free named entity disambiguation with wikipedia. Information. 2022;13(8):367.

[2] Szymański J, Olewniczak S, Piotrowski M, Pont MTS, Mora H. Entity Annotation with Wikipedia Using Neural Networks. In: Proceedings of the International Conference on Computer Information Systems and Industrial Management; 2024 Aug; Cham. Cham: Springer Nature Switzerland; 2024. p. 272-284.

[3] Nugues P. Linking Named Entities in Diderot's Encyclopédie to Wikidata. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation; 2024 May; Torino. 2024. p. 10610-10615.

[4] Wang H, Chowdhury SMH. Wikipedia Empowered Natural Language Interface for Web Search. In: Proceedings of the International Conference on Web Information Systems Engineering; 2024 Nov; Singapore. Singapore: Springer Nature Singapore; 2024. p. 14-25.

[5] Lippolis AN, Klironomos A, Milon-Flores DF, Zheng H, Jouglar A, Norouzi E, Hogan A. Enhancing entity alignment between wikidata and artgraph using llms. In: Proceedings of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage (SWODCH); 2023; Aachen: CEUR-WS; 2023. p. 1-12.

[6] Yan Z, Peng R, Wang Y, Li W. CTEA: Context and Topic Enhanced Entity Alignment for Knowledge Graphs. Neurocomputing. 2020;410:419-431.

[7] Heist N, Paulheim H. CaLiGraph: A knowledge Graph From Wikipedia categories and lists. Semantic Web. 2025;16(5):22104968251361349.

[8] Chen Y, Guo J, Hao C, Song C. Multi-Source Data Fusion Method Research on the Reconstruction and Expansion Project of Long-Line Expressway. Technical Gazette. 2025;32(1):149-156.

[9] Minardi S, Greco S, Barban N. A Comparison of Rule-based and Supervised Machine Learning Approaches for Record Linkage of Italian Historical Data. Historical Life Course Studies. 2025;15:28-46.

[10] Wang H, Zhao K, Li M, Zhang Y. Multi-modal Entity Alignment Based on Multidimensional Semantic Extraction. IEICE Transactions on Information and Systems. 2025;E108-D(5):2024EDP7173.

[11] Bai, Luyi, Song Xiuting and Lin Zhu. Joint multi-feature information entity alignment for cross-lingual temporal knowledge graph with bert. IEEE Transactions on Big Data 11.2 (2024): 345-358.

[12] Song, Hao, Yuxia Lei, Kangli Zi, Fuyuan Quan, Tianhang Zhang, and Qi Li. "An Entity Alignment Algorithm Based on Ontology Enhancement and Multi-Feature Fusion." In 2024 6th International Conference on Frontier Technologies of Information and Computer (ICFTIC), IEEE, 2024:942-948.

[13] Akhtar MU, Wang Y, Chen X, Li Z. Multilingual entity alignment by abductive knowledge reasoning on multiple knowledge graphs. Engineering Applications of Artificial Intelligence. 2025;139:109660.

[14] Zhang, Y., Chen, L., & Liu, Y. (2021). A Knowledge Graph Construction Method for Chinese Cultural Relics Based on Multi-source Data Fusion. Journal of Cultural Heritage, 48, 112–120.

[15] Li, H., & Wang, X. (2020). Entity Recognition and Alignment in Chinese Cultural Heritage Texts.

In Proceedings of the 2020 International Conference on Artificial Intelligence and Cultural Heritage (AICH 2020) (pp. 45–52). Springer.

[16] Dinges A, Hinze R. Truly Functional Solutions to the Longest Uptrend Problem (Functional Pearl). Proceedings of the ACM on Programming Languages. 2025;9(ICFP):463-478.

[17] Huang J, Li T, Jia Z, et al. Entity alignment of Chinese heterogeneous encyclopedia knowledge base[J]. Journal of computational and applied, 2016, 7(36):1881–1886.

[18] Singh PK, Singh KN. A similarity-based semi-supervised algorithm for labeling unlabeled text data. Expert Systems with Applications. 2025;258:128941.

[19] Zhao L, Wang B, Gao J, Li X, Hu Y, Yin B. Multi-modal Entity in One Word: Aligning Multi-level Semantics for Multi-modal Knowledge Graph Completion. IEEE Transactions on Big Data. 2025;12(3):1234-1248.

[20] Rifaldy F, Sibaroni Y, Prasetiyowati SS. Effectiveness of Word2VEC and TF-IDF in sentiment classification on online investment platforms using support Vector Machine. Jurnal Ilmiah Penelitian dan Pembelajaran Informatika. 2025;10(2):863-874.

[21] Chen, Y., Li, K., & Zhang, M. (2019). Enhancing Short Text Representation with External Knowledge and N-gram Features for Classification. *IEEE Access*, 7, 160122-160131.

[22] Taher HA, Hasan NNABM. Integration Named Entity Recognition and Latent Dirichlet Allocation to Enhance Topic Modeling. Annals of Emerging Technologies in Computing. 2025;9(2):45-56.

[23] Wang L, Jiao M, Li Z, Zhang M, Wei H, Liu Y. Image Captioning Model Based on Multi-Step Cross-Attention Cross-Modal Alignment and External Commonsense Knowledge Augmentation. Electronics. 2025;14(16):3325.

[24] Lan J, Qian X. Research on Improved RBM Recommendation Algorithm Based on Gibbs Sampling. Scalable Computing: Practice and Experience. 2025;26(3):1017-1034.