future internet

Article

Coarse-to-Fine Entity Alignment for Chinese Heterogeneous Encyclopedia Knowledge Base

Meng Wu [1], Tingting Jiang [1], Chenyang Bu [1,*] and Bin Zhu [2,*]

[1] Ministry of Education Key Laboratory of Knowledge Engineering with Big Data, School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, China; wu@mail.hfut.edu.cn (M.W.); jiangtt@mail.hfut.edu.cn (T.J.)

[2] Anhui Province Key Laboratory of Infrared and Low Temperature Plasma, College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China

* Correspondence: chenyangbu@hfut.edu.cn (C.B.); zhubin@nudt.edu.cn (B.Z.)

Abstract: Entity alignment (EA) aims to automatically determine whether an entity pair in different knowledge bases or knowledge graphs refer to the same entity in reality. Inspired by human cognitive mechanisms, we propose a coarse-to-fine entity alignment model (called CFEA) consisting of three stages: coarse-grained, middle-grained, and fine-grained. In the coarse-grained stage, a pruning strategy based on the restriction of entity types is adopted to reduce the number of candidate matching entities. The goal of this stage is to filter out pairs of entities that are clearly not the same entity. In the middle-grained stage, we calculate the similarity of entity pairs through some key attribute values and matched attribute values, the goal of which is to identify the entity pairs that are obviously not the same entity or are obviously the same entity. After this step, the number of candidate entity pairs is further reduced. In the fine-grained stage, contextual information, such as abstract and description text, is considered, and topic modeling is carried out to achieve more accurate matching. The basic idea of this stage is to use more information to help judge entity pairs that are difficult to distinguish using basic information from the first two stages. The experimental results on real-world datasets verify the effectiveness of our model compared with baselines.

Keywords: entity alignment; coarse-to-fine; Chinese knowledge base

check for updates

Citation: Wu, M.; Jiang, T.; Bu, C.; Zhu, B. Coarse-to-Fine Entity Alignment for Chinese Heterogeneous Encyclopedia Knowledge Base. *Future Internet* **2022**, *14*, 39. https://doi.org/10.3390/fi14020039

Academic Editor: Claude Chaudet

Received: 18 December 2021
Accepted: 20 January 2022
Published: 25 January 2022

1. Introduction

In recent years, researchers have made many efforts to obtain knowledge from the Web [1–5], and have constructed encyclopedic knowledge bases or knowledge graphs covering various fields [6]. Representative knowledge bases include YAGO [7], FreeBase [8], Dbpedia [9], Omega [10], etc. Chinese knowledge bases include Baidu Baike, Sogou Knowledge Cube, Fudan University Chinese General Encyclopedia Knowledge Graph (CN-Dbpedia) [11], Tsinghua University Bilingual Knowledge Base (Xlore) [12], etc. They are widely used in popular fields, such as search engines, question answering, recommendation systems, and natural language processing.

There is an urgent need to effectively integrate knowledge data from multiple heterogeneous data sources, and to finally form a large encyclopedic knowledge base with a clear logical structure and complete contents. Therefore, it is extremely significant to be able to associate various entities from different data sources with the same entity in the real world, that is, the task of entity alignment [13,14].

Current entity alignment methods are mainly divided into two categories [13,14], embedding-based methods and symbol-based methods. The advantage of embedding-based methods is that they are computationally efficient, because the similarity of two entities is calculated by the distance between their vectors. However, current embedding-based methods require a large amount of labeled data to train the model. The advantage of the symbol-based method is that it does not rely on label data and has high accuracy, but it has a high complexity and is time consuming, because the similarities of entity pairs composed of every two entities need to be calculated. Moreover, since the same entity may have different names, judging two entities based only on symbolic similarity will result in a low recall rate.

In this paper, we propose a three-stage model, inspired by the human brain's "large-scale-first" cognitive mechanism (from coarse-grained to fine-grained), for aligning heterogeneous Chinese encyclopedia knowledge bases. The model is named coarse-to-fine entity alignment (CFEA). The experimental results on real Chinese encyclopedia knowledge graph data show the effectiveness of our method. Our major contributions are as follows.

- We design a three-stage unsupervised entity alignment algorithm from the perspective of the human "large-scale-first" cognitive mechanism. It gradually prunes and matches candidate entity pairs from coarse-grained to fine-grained. We uses simple methods in the first and second stages, which greatly reduces the number of entity pairs that need to be compared, thereby improving the efficiency and accuracy of the algorithm through pruning.
- We combine different types of information to improve performance. That is, in three different stages of the model, three aspects of information are used, including entity type information, attribute information, and text information. Ordered by the difficulty in identifying entity pairs, the amount of information used and the complexity of the model gradually increase from the first stage through the third stage.
- We present an experimental study of a real Chinese encyclopedia dataset containing entity attributes and contexts. The experimental results demonstrate that our algorithm outperforms baseline methods.

The rest of the paper is structured as follows. Section 2 summarizes the related work. Section 3 explains our proposed CFEA model. Section 4 presents the experimental result and analysis. Section 5 concludes this paper.

2. Related Work

In this section, we first analyze CFEA and its related entity alignment models from two perspectives. We categorize them from the perspectives of informational scope and method of analysis, which are explained in Sections 2.1 and 2.2, respectively. After that, we analyze the relevant EA model for a Chinese encyclopedia knowledge base in Section 2.3.

2.1. Collective Alignment

From the perspective of the scope of information to be considered, EA methods can be divided into pairs-wise alignment, local collective alignment, and global collective alignment [15].

The paired-entity alignment method is a method of comparing the attribute information of the current matching entities in pairs. This method considers the similarity of entity attributes or specific attribute information, but does not consider the relationships between matching entities. Newcombe [16] and Fellegi [17] established a probabilistic model of the entity-matching problem, based on attribute similarity scores, by converting this problem into a classification problem,which is divided into matching, possible matching, and non-matching. This model is an important method of entity alignment. An intuitive classification method is to add the similarity scores of all attributes to obtain the total similarity score, and then set two similarity thresholds to determine which similarity interval the total similarity score sim_{attr}^{sum} is within. This can be expressed as:

$$\begin{cases} u \le sim_{attr}^{sum}(e_1, e_2) & \Rightarrow e_1, e_2 \text{ matched} \\ v \le sim_{attr}^{sum}(e_1, e_2) < u & \Rightarrow e_1, e_2 \text{ possible matched} \\ sim_{attr}^{sum}(e_1, e_2) < v & \Rightarrow e_1, e_2 \text{ mismatched} \end{cases} \tag{1}$$

Here, e_1, e_2 are the entity pairs to be matched; u and v are the upper and lower thresholds; sim_{attr}^{sum} firstly calculates matched pairs of attributes from e_1, e_2 respectively and then sums them to get a total similarity score. The main problem with this simple method is that it does not reflect the influence of different attributes on the final similarity. An important solution is to assign different weights to each matched attribute to reflect its importance to the alignment result. Herzog described this idea formally by establishing a probability-based entity link model based on the Fellegi–Sunter model [18]. The paired entity alignment method is simple to operate and convenient to calculate, but it has the disadvantage of considering less information and lacking semantic information. The middle-grained part in the CFEA model proposed in this paper is also constructed based on the classification idea and on the distribution weight idea, which belongs to the paired-entity alignment method.

Local collective entity alignment not only considers the attribute similarity of matching entity pairs, but also considers the similarity of their relationships. One method is to assign different weights to the attributes of the entity itself and the attributes of its related neighbor entities, and to calculate the overall similarity by weighted summation. This method can be formalized as:

$$sim(e_1, e_2) = \alpha sim_{attr}(e_1, e_2) + (1 - \alpha) sim_{NB}(e_1, e_2) \tag{2}$$

and

$$\begin{aligned} sim_{attr}(e_1, e_2) &= \textstyle\sum_{(a_1,a_2) \in attr(e_1,e_2)} sim(a_1, a_2); \\ sim_{NB}(e_1, e_2) &= \textstyle\sum_{(e_1',e_2') \in NB(e_1,e_2)} sim_{attr}(e_1', e_2'). \end{aligned} \tag{3}$$

where $sim_{attr}(e_1, e_2)$ is the attribute similarity function of the entity pair; $sim_{NB}(e_1, e_2)$ is the degree of neighbor similarity in the entity pair function; $0 \leq \alpha \leq 1$ is the adjustment parameter of the above two functions. This method takes the relationship of the entity as a special attribute of the entity and brings it into the calculation. In essence, it is still a paired-entity alignment method. The local collective entity alignment method does not really employ a "collective" approach.

The global collective entity alignment is a method of entity alignment that truly realizes the collective approach. One way is collective alignment based on similarity propagation. This method iteratively generates new matches [19–21] in a "bootstrapping" manner through initial matching. For example, two similarly named authors who have a "coauthor" relationship with two aligned author entities may have a higher degree of similarity, and then this similarity is propagated to more entities. The other way is based on a probability model. This type of method establishes a complex probability model for entities with which to match relationships and match decision-making. Generally, statistical relational learning is used for calculation and reasoning, and methods such as logical representation, probabilistic reasoning, uncertainty processing, machine learning, and data mining are integrated with relations to obtain the likelihood model in relational data. Bayesian network models [22,23], LDA models [24,25], CRF (conditional random field) models [26,27] and MLN (Markov logic networks) [28,29] models are commonly used probability models. Probabilistic models provide a standard method of relationship modeling, which can effectively improve the matching effort. Among them, the LDA model is an unsupervised model that mines the potential topic information of the text. It can simplify the data representation in a large-scale data set and retain basic information for data analysis, such as correlation, similarity, or clustering. It has the advantages of wide applicability and adaptability to large-scale data sets.

Among these probabilistic models, latent Dirichlet allocation (LDA) [30] is a three-layer Bayesian probabilistic model that consists of three layers: words, topics, and documents. It utilizes priori parameters α, β, the corpus document, D, and the parameters θ_d, ϕ_k to automatically generate the document and obtain the the posteriori $P(\theta, \phi|D)$ by learning the corpus document. The derived parameters θ_d, ϕ_k can determine the topic of the document at the semantic level. Figure 1 shows the Bayesian network diagram of the process of document generation with the LDA model. The corresponding process of document generation is as follows.

1. Sampling in the distribution $Dirichlet(\vec{\alpha})$ to get the topic distribution θ_d of document d.

2. Sampling in the distribution $Multinomial(\theta_d)$ to get the topic z_n of word n in document d.
3. Sampling in the distribution $Dirichlet(\vec{\beta})$ to get the word distribution ϕ_k of topic d (denoted as k).
4. Sampling in the distribution $Multinomial(\phi_k)$ to get the word w_n Finally, the document is generated after the above multiple sampling.

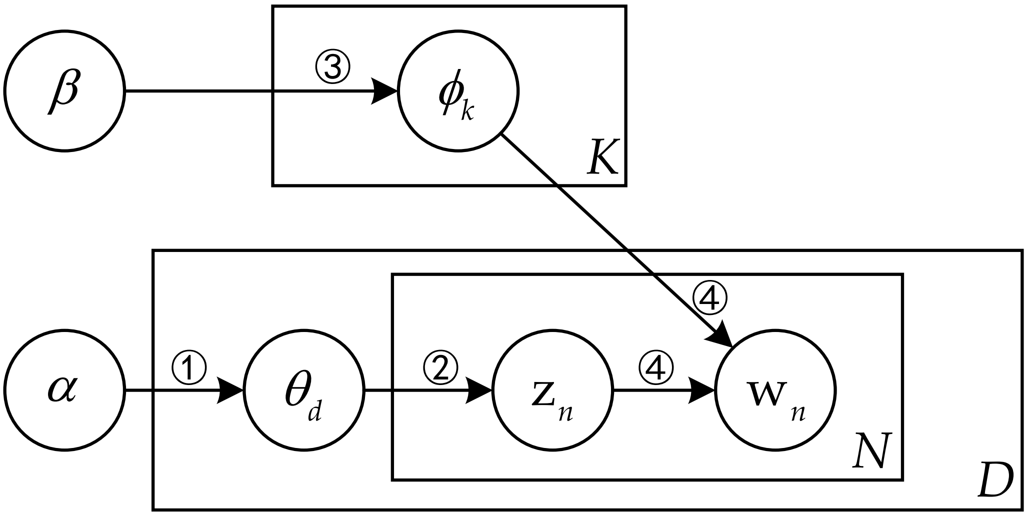

Figure 1. Bayesian Network for the LDA model. This diagram illustrates the process of generating documents. The symbol at the bottom right of the box represents the number of samples in the box, K represents the number of topics, N represents the number of words, and D represents the number of documents.

The CFEA (coarse-to-fine entity alignment) model proposed in this paper combines paired entity alignment and collective entity alignment. The middle-grained model (detail in Section 3.4) only considers the attribute information of the entity pair currently to be matched, thus, it is a paired-entity alignment method. The fine-grained model (detail in Section 3.5) mainly uses the LDA method. The LDA topic model obtains additional entity information that cannot be obtained from attribute information from the semantic level of the text to help align entities. This is a global collective entity alignment method. It performs topic modeling on all texts in the data set, and there is a topic-level influence between any two entities.

2.2. Similarity-Based Alignment

From the perspective of information analysis methods, entity alignment methods could be classified as follows: based on network ontology semantics [31], based on rule analysis and based on similarity theory judgment.

Entity alignment based on web ontology semantics uses the web ontology language (OWL) to describe the ontology and to perform reasoning and alignment. This type of method combines the semantic information of OWL in the knowledge base, such as the heuristic reasoning algorithm [32], iterative discriminant model [33], to reason and judge whether it can be aligned. However, for Chinese encyclopedia knowledge bases, such as Baidu Baike and Hudong Baike, do not contain complete ontology information, so it is difficult to align them according to OWL semantics.

The entity alignment method based on rule analysis uses the rule-based evaluation function to judge and align by formulating special rules in specific application scenarios. However, the method based on rule analysis is not universal. It is possible to perform alignment based on rule analysis only by establishing a large number of different rules in many specific areas. Similarly, Chinese Web Encyclopedia contains a large amount of entity neighborhood information, and it is difficult to align it by establishing a large number of different rules for specific fields.

Therefore, in the face of Chinese encyclopedia knowledge bases that do not have complete ontology information and that contain a large number of subject areas, more research is focused on the method of similarity theory determination. The method of judging based on the similarity theory is widely used. This method considers the attribute value information or description text information of the entity, and calculates the similarity of such information for alignment. Some research has used similarity-based fuzzy matching, topic models, etc. to align the entities of the encyclopedia and knowledge base. Among the methods of attribute similarity calculation, there are mainly token-based similarity calculations, LCS-based attribute similarity calculations, and edit distance-based similarity calculations.

- Token-based similarity calculation. This method uses a certain function to convert the text string to be matched into a collection of substrings, called collections of tokens, and then uses the similarity function to calculate the similarity of the two token sets to be matched, such as the Jaccard similarity function, cosine similarity function, and the qgram similarity function, etc. Different similarity functions have different characteristics.
- The LCS-based similarity calculation. This method finds the longest common subsequence in two text sequences, and the length ratio of the original sequence is used as the similarity.
- Edit distance-based similarity calculation. This method regards the text string to be matched as a whole, and the minimum cost of the editing operation required to convert one string to another string is used as a measure of the similarity of two strings. Basic editing operations include inserting, deleting, replacing, and swapping positions.

The middle-grained model of the CFEA model we propose uses similarity calculation based on edit distance. The similarity calculation based on edit distance can effectively deal with sensitive issues such as input errors, which is of great significance to the network encyclopedia data that belongs to the user's original content UGC. Commonly used similarity functions based on edit distance are based on Levenshtein distance [34], Smith–Waterman distance [35], affine gap distance [36], or Jaro and Jaro–Winkler distance [37,38]. The middle-grained model will use the based on Levenshtein distance similarity function, which will be introduced in detail in Section 3.4.

2.3. Alignment for Chinese Encyclopedia

The main method of existing Chinese encyclopedia knowledge base entity alignment is based on the calculation of attribute information similarity. The paper [39,40] only uses the attribute value of the entity to calculate the similarity of the entity and determine whether the entity can be aligned. The comprehensive index is poor due to the irregularity of attribute names and attribute values in the web encyclopedia.

In addition, the similarity calculation of contextual information composed of entity abstracts and descriptive texts is also an important method in the alignment of encyclopedia knowledge base entities. On the one hand, it is necessary to use additional text information as a supplement because of the heterogeneity and ambiguity of structured attribute information in different knowledge bases; on the other hand, using a large number of unstructured entity abstracts or descriptive text information in the knowledge base, we can obtain latent semantics and relationships and construct entity features, which can be helpful in further alignment. One way to consider contextual information is to use TF-IDF [39] for text modeling. However, traditional text modeling methods, such as TF-IDF, only consider the characteristics of word frequency, and do not consider the semantic relationship between terms. The LDA topic model can mine the underlying topic semantics of the text unsupervisedly. The model in [41] combined LCS and LDA, and achieved a high accuracy rate. However, the model lacked the granularity of a classification/partition index, and entailed no coarse-to-fine process. The text co-training model [42] use the semi-supervised method to learn the multiple features of entities. The paper [43] improved the LDA model, the comprehensive index improved under some datasets. However, it is difficult for a single LDA model to achieve a high accuracy rate on a wider dataset, and the time consumption of LDA similarity calculation and matching is relatively high. WA-Word2Vec [44] use LTP (language technology platform) to identity the named entities in the text. Then, they deal word vectors and obtain feature vectors by Word2Vec. The latest method is MED-Doc2Vec [45]. It used minimum edit distance and the Doc2vec model to obtain feature vectors containing semantic information. This model obtained comprehensive entity similarity by weighted average to complete the entity-alignment task and achieved better results in the experiment.

In summary, the entity alignment method currently applied to the Chinese Encyclopedia knowledge base is struggles to achieve a balance between accuracy and time. Therefore, we propose combining the edit distance-based method with the LDA topic modeling method, and apply the coarse-to-fine idea to the unsupervised Chinese encyclopedia entity-alignment task.

3. CFEA Model

In this section, we discuss some of the definitions and task descriptions in this paper in Section 3.1. Then, the framework of our model is described in Section 3.2. Finally, the three stages of the CFEA model are detailed in Sections 3.3–3.5, respectively.

3.1. Preliminaries and Task Description

In this subsection, we first discuss the definition of a heterogeneous encyclopedia knowledge base with the task description of entity alignment. Then, the notations used are listed.

Definition 1. *Heterogeneous encyclopedia knowledge base. The heterogeneous encyclopedia knowledge base comes from encyclopedia websites and contains the semi-structured attribute of entities and unstructured contextual information. Among them, the latter consists of abstracts and descriptive text. These types of information are diverse in structure and ambiguous in content in different knowledge bases. For instance, Baidu Baike, Hudong Baike, and Chinese Wikipedia are quite distinct in the structure and content of web pages.*

Definition 2. ***Entity Alignment.*** *Entity Alignment (EA, also called entity matching) aims to identify entities located in different knowledge bases that refer to the same real-world object. The illustration of entity alignment for knowledge base is shown in Figure 2.*

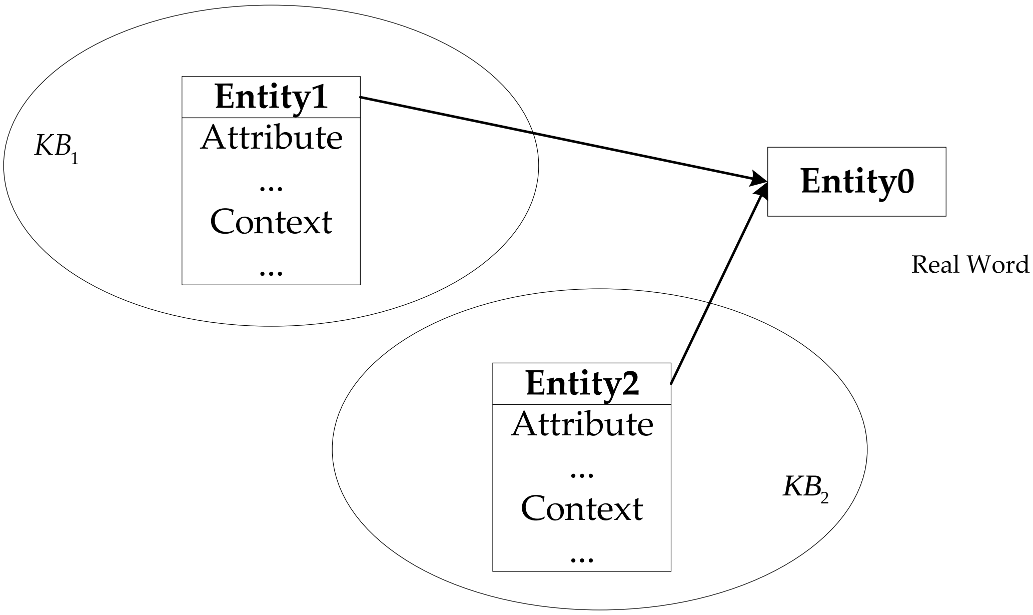

Figure 2. Illustration of knowledge base entity alignment (reproduced from [46]). KB_1 and KB_2 are two different knowledge bases, and Entity1 and Entity2 are the entities in the two knowledge bases, respectively. Both entities refer to the same entity Entity0 in the real world.

The notations used in this paper are listed in Table 1 in order of their appearance.

Table 1. Notation Overview.

Notation	Description
e	entity *e*
KB	knowledge base
E	set of knowledge base dataset
AE	aligned set of knowledge base dataset
a	weight of URL in entity's attributes
c	cutoff threshold of matched attributes
u	upper threshold in attributes match
v	lower threshold in attributes match
ω	threshold in LDA model
K	number of topics
N	number of words
D	number of documents
α	a priori parameter of the LDA model
β	a priori parameter of the LDA model
z_n	a topic of word *n*
w_n	a word with index *n*
θ_d	documents-topics matrix for document *d*
ϕ_k	topics–words matrix for topic *k*
$LevD(str_a, str_b)$	Levenshtein distance of string a and string b
W_e	set of feature words of entity *e*
$feature_e$	feature probability matrix of entity *e*
V_e	feature vector of entity *e*

3.2. Overview of CFEA

Figure 3 shows the framework of the CFEA model we designed.

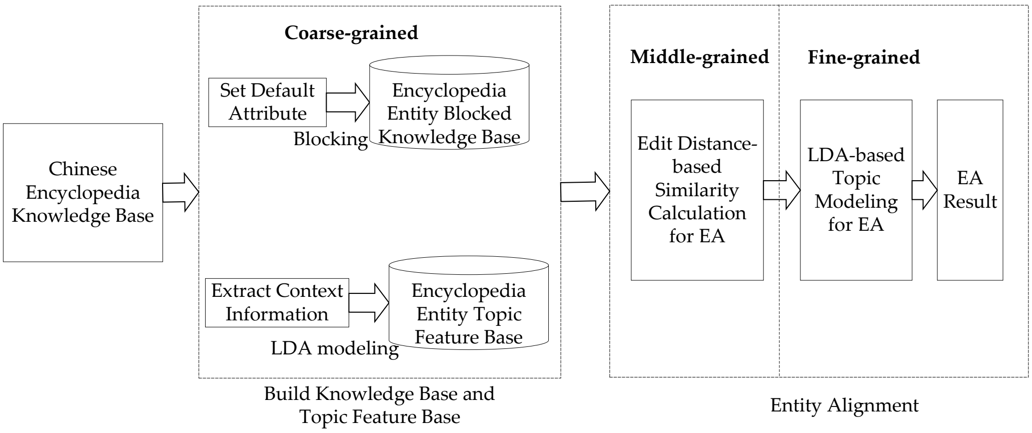
Chinese Encyclopedia Knowledge Base
Coarse-grained
Set Default Attribute
Blocking
Encyclopedia Entity Blocked Knowledge Base
Extract Context Information
LDA modeling
Encyclopedia Entity Topic Feature Base
Build Knowledge Base and Topic Feature Base
Middle-grained
Edit Distance-based Similarity Calculation for EA
Fine-grained
LDA-based Topic Modeling for EA
EA Result
Entity Alignment

Figure 3. Framework of the CFEA model.

The steps of the CFEA model as applied to a heterogeneous encyclopedia knowledge base are as follows. (1) We first extract the attributes, abstracts, and descriptions of the entities in the dataset, and combine the latter two as contextual information. (2) Secondly, we block the knowledge base and build the topic feature base. The blocking algorithm refers to grouping entities that meet the attribute requirements into the corresponding categories. It is the coarse-grained part of the CFEA. The construction of the topic feature base requires all contexts to be processed by Chinese word separation and input to the LDA model for unsupervised training, which, in turn, yields the latent topic features of the entity. (3) After that, it is the process of aligning the entities. For pairs of entities to be aligned from different data sources, we use the method of the edit distance-based similarity calculation of attributes as the middle-grained part for preliminary judgment. If the similarity score meets the upper boundary of the threshold, the pair of entities is directly judged as aligned. If beneath the lower threshold, pairs are determined as non-aligned. Otherwise, it is between the upper and lower threshold, which means that the current attribute information is insufficient for judgment. (4) At last, we explore the fine-grained perspective, with the well-trained LDA topic model. We further consider the contextual information of the current entities and perform the calculation of topic similarity. After the above steps, all matching entities are generated in a coarse-to-fine process.

The process of coarse-to-fine entity alignment, described in Algorithm 1.

Human cognition is a process of refinement from coarse-grained perspectives to a fine-grained perspectives [47]. Most existing collective entity alignment methods have high time complexity. Therefore, we introduce the human brain's "global precedence" cognitive mechanism. We first obtain the category to which each entity belongs as the coarse-grained step and weed out candidates with different types in the following process of entity alignment. As these parts of the entities cannot reach a match with other types of entities, in this way, our approach achieves a reduction in runtime by reducing the size of the candidate knowledge base.

3.3. Coarse-Grained Blocking Algorithm

For instance, the search for the keyword "Milan" in Chinese on the website of Baidu Baike yields 21 meanings. As in Figure 4, the diagram shows some of the meanings, such as a city in Italy, the Milan soccer club, and an orchid plant. These meanings can be classified as city names, names of people, names of books or films, names of plants, names of sports clubs, etc. When the current entity named "Milan" is judged by its attributes to be a person's name rather than a city's name or something else, we only need to look for candidates among the category of person entities. As a result, the number of candidate entities is significantly reduced from 21 to a single-digit number.

Algorithm 1: Coarse-to-fine entity alignment algorithm

```
Input: dataset $E_A$,$E_B$, weight $a$, thresholds $c$, $u$, $v$, $\omega$, number of topics $K$
Output: result of entity alignment $AE$
foreach entity $e \in (E_A \cup E_B)$ do
    build LDA model;
    compute topics–words matrix $\phi$ and documents–topics matrix $\theta$ ;
    compute $e_{type}$ and partition ;
end
for $i \leftarrow 1$ to $size(E_A)$ do
    for $j \leftarrow 1$ to $size(E_B)$ do
        if $i, j$ in same partition then
            compute propertySim=EditDist($e_i, e_j$) ;
            // by edit distance-based similarity
            if $propertySim > threshold\ u$ then
                $AE \leftarrow AE \cup \{(e_i, e_j)\}$;
            else if $propertySim < threshold\ v$ then
                continue;
            else
                compute contextSim=LDASim($e_i, e_j$) ;
                // by LDA topic feature similarity
                if $contextSim > threshold\ \omega$ then
                    $AE \leftarrow AE \cup \{(e_i, e_j)\}$;
    end
end
```

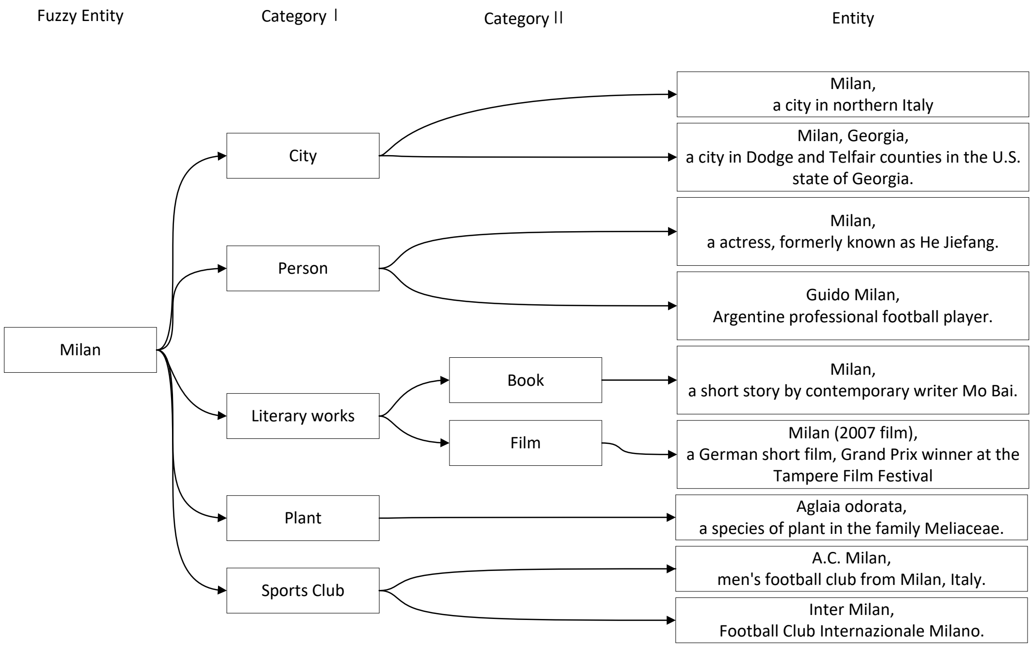
Fuzzy Entity
Category Ⅰ
Category Ⅱ
Entity
Milan
City
Person
Literary works
Plant
Sports Club
Book
Film
Milan,
a city in northern Italy
Milan, Georgia,
a city in Dodge and Telfair counties in the U.S. state of Georgia.
Milan,
a actress, formerly known as He Jiefang.
Guido Milan,
Argentine professional football player.
Milan,
a short story by contemporary writer Mo Bai.
Milan (2007 film),
a German short film, Grand Prix winner at the Tampere Film Festival
Aglaia odorata,
a species of plant in the family Meliaceae.
A.C. Milan,
men's football club from Milan, Italy.
Inter Milan,
Football Club Internazionale Milano.

Figure 4. An example of the entity "Milan" in Baidu Baike. The multiple meanings of Milan are displayed in the rightmost column in the figure. These meanings can be divided into several categories, which are shown in the middle of the figure.

In the field of database technology, the method of pre-processing before entity matching is called "blocking and index techniques" [48]. Blocking and index techniques assign entities with certain similar characteristics to the same block, so that entity matching is done only within the same block. This technique, which comes from the database domain, can be applied to the entity alignment for the knowledge base.

Our method requires setting a "preset attribute list". As shown in Figure 5a, it is a two-dimensional list, in which the first dimension contains several categories that we include in the reference and the second dimension contains some names of feature attributes we set for that category.

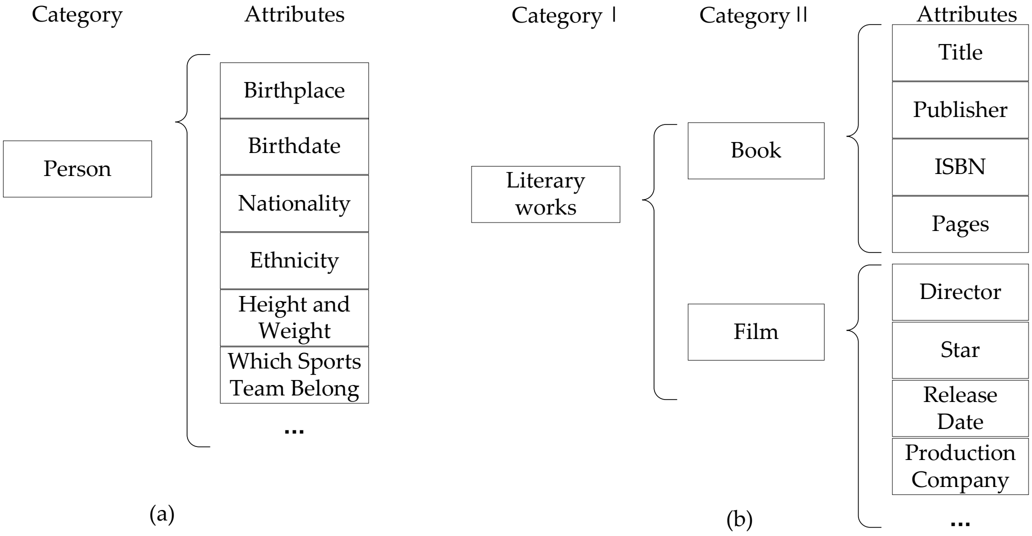

Figure 5. Examples of the preset attribute list. (**a**) is a category in a two-dimensional list, while (**b**) is a coarse category in a three-dimensional list. Each category corresponds to some attributes, which are called preset attributes. Set categories' preset attributes are based on experience to distinguish the types of entities at multiple granularities.

The preset attribute list is versatile and flexible for blocking the encyclopedic knowledge base, which usually contains many kinds of entities. Based on the official category indexes of several encyclopedia sites, and removing categories with little difference or containment–inclusion relationships, we get the preset attribute list after collating them. This is generally applicable to other Chinese encyclopedic knowledge bases as well.

In addition, the list can be extended somewhat to improve the efficiency and accuracy of blocking. If the two-dimensional list is extended to a multi-dimensional list, i.e., there are small categories in a large classification, and the scope and accuracy of the classification is flexibly controlled at multiple levels. Figure 5b is an example of coarse-grained category in a three-dimensional list. It allows more flexibility to control whether the matching range is coarse or fine, and whether the entity matching speed is fast or slow, and to find a balance when performing matching between entities.

3.4. Middle-Grained Attribute Similarity

The attributes of entities in the knowledge base are the fundamental information for the study of alignment. However, there are two types of problems with the attributes in the Chinese web encyclopedia: on the one hand, the heterogeneous encyclopedic knowledge base contains voluminous contents of subject areas and does not have complete ontological information; on the other hand, the web data of user-generated content (UGC) introduces the problem of ambiguity.

For the former problem, the similarity-based approach allows distinguishing the entities in the encyclopedic knowledge base, especially those with more accurate attribute definitions. For the latter problem, we use an edit distance-based approach to calculate the similarity. The reason for this is that an edit distance-based approach can effectively deal with error sensitivity issues, such as entry errors. This is meaningful for web encyclopedia data that belongs to UGC. Therefore, we propose an attribute similarity calculation method suitable for encyclopedic knowledge bases based on the edit distance algorithm described in [34].

Specifically, the Levenshtein distance is a typical method of edit distance calculation that measures the similarity of strings. Edit distance is the minimum number of operations to convert the string *str*1 to *str*2. The operations for characters include insertion, deletion, and replacement. The smaller the distance, the more similar the two strings are. For two strings S1 and S2, the Levenshtein distance calculation formula [34] is as follows.

$$LevD_{S1,S2}(i,j) = \begin{cases} i & ,\text{if } j = 0 \\ j & ,\text{if } i = 0 \\ min\begin{cases} LevD_{S1,S2}(i-1,j)+1 \\ LevD_{S1,S2}(i,j-1)+1 \\ LevD_{S1,S2}(i-1,j-1)+1 \quad {}_{S1[i]\neq S2[j]} \end{cases} & ,otherwise \end{cases} \tag{4}$$

where $S1[i] and S2[j]$ are the i th character of string S1 and the j th character of string S2, respectively. From the above, we apply edit distance to attribute name alignment and attribute value similarity calculation, respectively. Different knowledge bases vary not only in describing textual information and semi-structured attribute's values, but also in what attributes they contain and which attribute's names they call. The first use of edit distance-based similarity calculation in our CFEA model is to match the attribute's names of different entities, i.e., attribute alignment. Only attributes that reach a certain threshold will be matched for the next values alignment, otherwise they will be treated as invalid information. The second use of edit distance calculation is to perform a similarity calculation on the attribute values of the entity pairs to be matched. The detailed calculation is shown below.

1. Attributes of encyclopedia knowledge base

   Definition 3. *After the preprocessing of the dataset, entity e_a has the set of attribute names, that is, $Attribute_a = p_{a1}, p_{a2}, p_{a3}, ..., p_{am}$. The corresponding set of attribute values is $Value_a = v_{a1}, v_{a2}, v_{a3}, ..., v_{am}$. Entity e_b's set of attribute names is $Attribute_b = p_{b1}, p_{b2}, p_{b3}, ..., p_{bn}$. Its set of attribute's values is $Value_b = v_{b1}, v_{b2}, v_{b3}, ..., v_{bm}$. where m, n are the number of attributes of the entity e_a, e_b, respectively.*

2. The conditions for matching attribute $Attribute_a$ and $Attribute_b$

   $$LevD(p_a, p_b) > c \tag{5}$$

   where p_a and p_b are the attribute's names from $Attribute_a$ and $Attribute_b$, respectively, and c is the minimum threshold for an attribute name to match.
3. For the matched attribute names $p_i \in interAttribute(e_a, e_b)$, the similarity of attribute p_i is calculated by

   $$sim(p_i) = \frac{LevD(v_{ai}, v_{bi})}{max(len(v_{ai}), len(v_{bi}))} \tag{6}$$

   where v_{ai}, v_{bi} are the attribute values corresponding to the attribute names p_i in $Value_a and Value_b$, respectively.
4. The attribute similarities between the entities e_a and e_b are calculated as

   $$AttributeSim(e_a, e_b) = \left[\sum_{i=1}^{T} sim(p_i)\right] / T \tag{7}$$

   where $T = len(interAttribute(e_a, e_b))$

3.5. Fine-Grained Context Topicalization

The effectiveness of simple attribute similarity-based matching is insufficient, especially in the case of missing and ambiguous information of attributes. However, most of the Chinese web encyclopedia websites have the problems of heterogeneity and ambiguity, such as the lack of attributes and the complicated interleaving of information. This causes difficulties for the entity-matching problem based on attribute information only. Therefore, additional contextual information is needed to supplement it.

Entities in an encyclopedic knowledge base site not only contain structured or semi-structured knowledge such as attribute tables but usually also contain texts of description and abstracts of the webpages to which it belongs. The description text describes the entity from different aspects, and the abstract provides a brief summary of the relevant webpages. We refer to the two together as the context information of an entity. Using the vast amount of unstructured contexts present in the knowledge base, we obtain the underlying semantics and relationships and construct entity features for deeper studies through complex probabilistic model building and statistical relationship learning.

An essential feature is that the contexts for the same entity may differ across different knowledge bases, but their topics are similar. Therefore, to exclude the interference caused by textual differences, we perform LDA-based topic modeling of the context and apply it to fine-grained entity alignment.

The similarity calculation of contexts using the LDA topic model includes the following steps:

1. Data acquisition and pre-processing
   In this part, we obtained corpus from Baidu Baike and Hudong Baike. These corpora include abstracts and descriptions of encyclopedic entities. After that, we perform word splitting for Chinese and filter out the stopwords. We sorted out four stop-word lists (https://github.com/goto456/stopwords accessed on 17 December 2021), cn-stopwords, hit-stopwords, baidu-stopwords, and scu-stopwords from different companies or universities with our supplements.
2. Topic modeling and parameter estimation
   LDA modeling finds the topic distribution of each document and the word distribution of each topic.
   LDA assumes that the prior document–topic distribution is the Dirichlet distribution, that is, for any document d, its topic distribution θ_d is:

$$\theta_d = Dirichlet(\vec{\alpha}) \tag{8}$$

   where α is the hyperparameter of the distribution and is a K-dimensional vector, and K is an artificially pre-defined parameter for the number of topics.
   Furthermore, LDA assumes that the prior topic–word distribution is also the Dirichlet distribution, that is, for any topic k, its word distribution ϕ_k is:

$$\phi_k = Dirichlet(\vec{\beta}) \tag{9}$$

   where β is the hyperparameter of the distribution and is a V-dimensional vector, and V is the number of words.
   Based on the Dirichlet distribution, we conduct topic modeling on the corpus after pre-processing. Then perform parameter estimation for topic model, that is, solve θ_d, ϕ_k with Gibbs sampling [49]. After that, we can generate the topic feature with solved parameters.
3. Topic feature generation
   Topic-word matrix ϕ_k is as follows.

$$\phi_k = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1v} \\ p_{21} & p_{22} & \cdots & p_{2v} \\ \vdots & \vdots & & \vdots \\ p_{k1} & p_{k2} & \cdots & p_{kv} \end{bmatrix} \tag{10}$$

   where p_{ij} represents the probability of assigning the j th word to the i th topic, v is the total number of words, and k is the total number of topics.

We take the m ($m < v$) words with the highest probability in each row of records $(p_{i1}, p_{i2}, \cdots, p_{iv})$ in the topic–word matrix ϕ_k. That is, the set of feature words and the feature matrix are obtained by taking m words under K topics, as shown in Equations (11) and (12), respectively.

$$W_e = \{w'_1, w'_2, \cdots, w'_n\} \tag{11}$$

where n is the number of non-repeating feature words ($n \leq K \cdot m$).

$$feature_e = \begin{bmatrix} p'_{11} & p'_{12} & \cdots & p'_{1m} \\ p'_{21} & p'_{22} & \cdots & p'_{2m} \\ \vdots & \vdots & & \vdots \\ p'_{k1} & p'_{k2} & \cdots & p'_{km} \end{bmatrix} \tag{12}$$

4. Similarity calculation
For each word $w'_i \quad i \in (1, 2, \cdots, n)$ in the set of feature words W_e, find its maximum value in the feature matrix $feature_e$ as the eigenvalue v'_i of the word. That means the feature vector is obtained.

$$V_e = (v'_0, v'_1, \cdots, v'_n) \tag{13}$$

For the entities $e_a and e_b$, their feature vectors are $V_{e_a} and V_{e_b}$, respectively. We use cosine similarity to process the feature vectors of different entities to get the similarity between contexts. The contextual similarity of the entity e_a, e_b is calculated as

$$contextSim(e_a, e_b) = \frac{V_{e_a} \cdot V_{e_b}}{|V_{e_a}||V_{e_b}|} \tag{14}$$

4. Experiment

To verify the performance of our CFEA model, we conducted a thorough experiment. We illustrate the dataset and experimental setting in Sections 4.1 and 4.2 respectively. Then, the evaluation metrics are described in Section 4.3. Finally, we discuss the performance of CFEA in Section 4.4.

4.1. Dataset

We download the Zhishi.me (http://openkg.cn/dataset/zhishi-me-dump accessed on 17 December 2021) which is a real dataset collection for a Chinese encyclopedia knowledge base from openkg.cn accessed on 17 December 2021. Due to the large size of the dataset, we extract part of the data from which the original source is Wikipedia (Chinese version), Baidu Baike or Hudong Baike. The extracted information includes entities, attributes, context, etc., contains approximately 300k entities, 2000k pieces of attributes, and 280k pieces of context. The basic statistics of the datasets are presented in Table 2.

Table 2. Basic statistics of the datasets.

	Baidu–Hudong		**Wikipedia–Baidu**	
entity	29,764	170,260	11,975	87,642
attribute	277,945	1,691,591	108,400	841,495
context	27,072	152,713	11,965	78,966
aligned entity	12,511		314	

4.2. Experiment Setting

To dig out the entity pairs that point to the same real-world entities in Wikipedia, Baidu Baike and Hudong Baike, we utilize the model to perform experiments that judge this automatically. According to the conventions of related experiments, we take out the entities from the two encyclopedias one by one and combine them into entity pairs. These entity pairs include all possible matching entity pairs, as well as a large number of unmatched entity pairs. Then, the entity pairs are processed through models whose direct task is to determine whether the entity pair points to the same real entity, pair by pair. The judgment result of the alignment is recorded and compared with the answer file to calculate the precision rate, recall rate, and other indicators.

We also reproduce several baselines models (introduced later) and test them to explore their respective effectiveness. We process the dataset in the python 3.8 environment and use the relevant NLP library such as Gensim. For Zhishi.me, the parameter settings of the CFEA model are shown in Table 3. Among them, value1 is for the data set Baidu–Hudong, and value2 is for Wikipedia–Baidu. The meanings of every parameter are listed in Table 1.

Table 3. The parameter settings of CFEA in our experiment.

Parameter	Value1	Value2
a	0.5	0.5
c	40	40
u	0.7	0.5
v	0.4	0.2
ω	0.7	0.7
K	14	16

The baselines for comparative experiments are as follows:

- ED (edit distance): Matching attributes of entities based on edit distance and aligning them based on similarity.
- Weighted ED: Weighting of attributes, especially the URL of the entity as an important aspect. Then, alignment experiments are conducted based on the weighted entity attributes.
- ED-TFIDF: The two-stage entity alignment model, similar to CFEA. First, it matches attribute information based on the weighted edit distance. Then, it calculates the TF-IDF value of each word in the context and uses the TF-IDF values of all words as the feature vector of that context. This is a text modeling approach that considers frequency features of words but has no semantic information.
- LDA [24,43]: This baseline uses LDA to model the context of entities and calculates the similarities between entities in different knowledge bases based on this. Bhattacharya et al. [24] used the LDA model in entity resolution. We apply LDA as a baseline and run it in dataset of this paper. In addition, the literature [43] has made certain improvements to the LDA model and applied it to the entity alignment task of the encyclopedia knowledge base.
- LCS [39]: This uses the entity's attributes to calculate the similarity based on LCS and determine whether the entity can be aligned or not.
- Weighted LCS [40]: This performs entity alignment based on LCS after weighting the attributes.
- LCS-TFIDF: A two-stage entity alignment model that uses TF-IDF to model the contexts in the second stage.
- LCS-LDA [41]: This is also a two-stage method of entity alignment; using LCS and LDA has been effective on its dataset.
- WA-Word2Vec [44]: An entity alignment method with weighted average Word2Vec. It uses LTP (language technology platform) to identity the named entities in the text. Then, it deals word vectors and feature vectors using Word2Vec. Finally, the cosine similarity is calculated to align the entities.
- MED-Doc2Vec [45]: A multi-information weighted fusion entity alignment algorithm. It uses minimum edit distance and the Doc2vec model to obtain feature vectors

containing semantic information; it obtains the entity's comprehensive similarity by weighted average to complete the entity-alignment task.

4.3. Evaluation Metrics

We use precision, recall, and F_1 score as evaluation metrics.

1. Precision

$$P = N_T/(N_T + N_F) \tag{15}$$

where N_T is the number of correctly aligned entity pairs in the experimental results, while N_F is the number of incorrectly aligned entity pairs in the experimental results. Precision indicates the fraction of positive instances among the retrieved instances.

2. Recall

$$R = N_T/N_A \tag{16}$$

where N_A is the number of all alignable entities in the dataset. Recall is the fraction of relevant instances that were retrieved.

3. F_1 score

$$F_1 = 2PR/(P + R) \tag{17}$$

F_1 is a composite measure of precision and recall.

4.4. Experiment Results

In order to verify the effectiveness of the CFEA model, we conducted comparison experiments using several baselines. Moreover, we explored the effect of the number of topics parameter, i.e., the parameter K and LDA's threshold ω on the model using CFEA with different values. The experimental results and analysis are shown below.

Table 4 shows that the proposed method obtains competitive results on both datasets. For example, for the Baidu–Hudong dataset, our method outperforms other compared algorithms on both the F1 metric and the recall metric. As for the WiKi–Baidu dataset, our method performs the second best. The experimental results demonstrate the effectiveness of the proposed method. Moreover, Table 4 also shows that the two-stage models perform better than the single-stage models in most cases. This might be because more information (i.e., abstract and description) is used in these two-stage models.

Table 4. The results of F_1, precision and recall in entity alignment.

		Baidu–Hudong			**WiKi–Baidu**		
		F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
single-stage methods	ED	0.8815	0.9695	0.8082	0.6039	0.5087	0.7429
	weight ED	0.9143	0.9714	0.8635	0.7581	0.7705	0.7460
	LDA	0.7637	0.7330	0.7971	0.4449	0.3425	0.6349
	LCS	0.9020	0.9716	0.8418	0.7199	0.8153	0.6444
	weight LCS	0.9062	**0.9726**	0.8482	0.7542	0.7265	0.7841
two-stage methods	weight ED+TFIDF	0.9409	0.9697	0.9138	0.7586	0.7857	0.7333
	weight LCS+TFIDF	0.9388	0.9687	0.9106	0.7633	0.7147	**0.8190**
	weight LCS+LDA	0.9432	0.9669	0.9206	0.8006	0.8045	0.7968
	WA-Word2Vec	0.8062	0.7437	0.8801	0.6590	0.8382	0.5429
	MED-Doc2vec	0.9464	0.9691	0.9248	**0.8188**	**0.9073**	0.7460
proposed method	CFEA	**0.9472**	0.9680	**0.9273**	0.8099	0.8448	0.7778

To analyze the influence of the number of topics on the model, we experimentally explored the change curve of the related results for different values of the parameter K. The experimental results are shown in Figure 6. When K is set to 14, the model obtains the best F_1 in Baidu–Hudong. When K is less than 14, the experimental results show that the precision is poor, the recall rate is high, but F_1 is obviously poor. This might be because there are too few topics, and a topic will contain multiple layers of semantic information. At this time, the accuracy rate of the LDA model is low, because in the LDA model it is easy to misunderstand that some entity pairs have similar semantics in text but do not actually match, and so are matched anyway. When the K value is large, the accuracy of the model gradually increases, but the recall rate drops sharply. This might be because too many topics without obvious semantic information will make the matching of the topic model more stringent, that is, only very similar texts can be matched, and the rest will be discarded. The dataset Wikipedia–Baidu has a similar pattern, and the only difference is that the optimal K is set to 16.

The impact of the similarity threshold of LDA on the overall model is given as follows. We change the threshold under the optimal parameter setting. Only the similarity scores that reach this threshold will be regarded as aligned, otherwise it will be judged as misaligned. The relevant experimental results are shown in Figure 7. The CFEA model obtains the best F_1 score when the parameter ω is 0.7. When the parameter is gradually increased from 0.1, the precision of the model gradually increases, but the recall is greatly reduced. The reason may be that a relatively high threshold will cause the model to be too strict on the topical similarity of the text. Although the precision has been improved, it leads to a sharp drop in the recall rate. This makes the F_1 score experience an increase, and then, gradually, a decrease after reaching the peak. In brief, the setting of the LDA threshold is extremely sensitive to the balance between precision and recall. Therefore, we should set a suitable threshold to strike a balance.

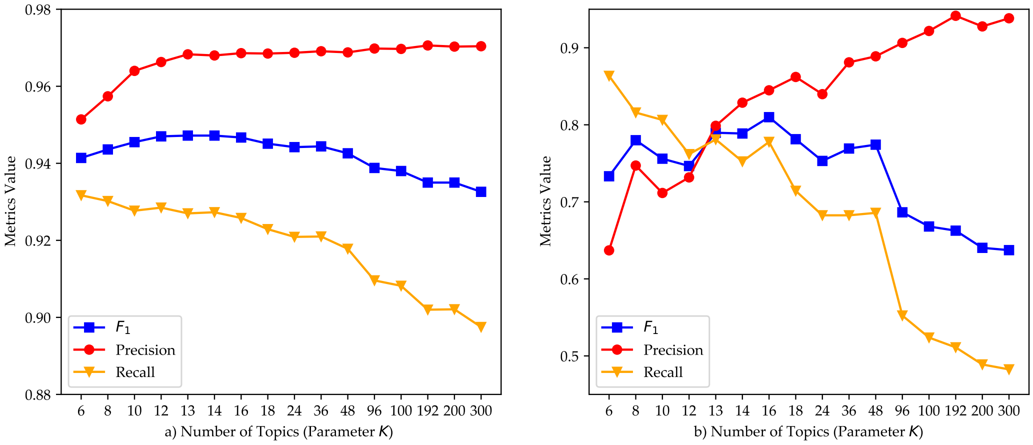

Figure 6. Comparison of entity alignment results of different number of topics; (**a**) Baidu–Hudong; (**b**) Wikipedia–Baidu.

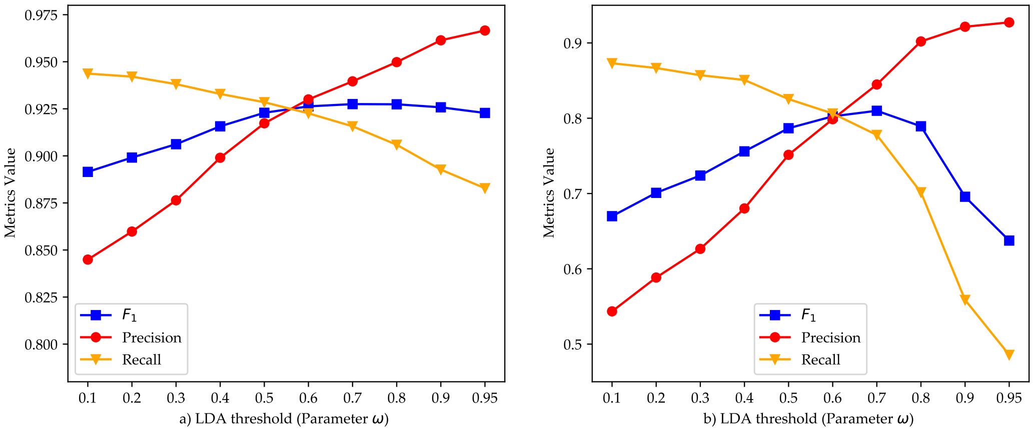

Figure 7. Comparison of entity alignment results of different LDA thresholds; (**a**) Baidu–Hudong; (**b**) Wikipedia–Baidu.

5. Conclusions and Future Work

In this paper, we proposed a coarse-to-fine entity alignment model for Chinese heterogeneous encyclopedia knowledge bases with three stages; a coarse-to-fine process combining the advantages of pruning strategy, attribute similarity-based methods, and context modeling methods for entity alignment. The experimental results with real-world datasets showed that our proposed model is better than several other algorithms.

In future research, the following topics are worthy of further study. Firstly, the blocking-based pruning algorithm needs further improvement. It requires too much manual participation for a blocking algorithm based on the multi-granularity preset attributes proposed in this paper. a blocking algorithm with less manual participation is a topic worthy of study. What is more, the LDA is a bag-of-words model, which does not consider the order of words and cannot dig deeper into semantic information. Therefore, the effectiveness of context modeling has room for further improvement in this field.

Author Contributions: Conceptualization, M.W.; Writing—original draft, M.W.; Writing—review & editing, T.J., C.B. and B.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of China (No. 61806065 and No. 62120106008), the Fundamental Research Funds for the Central Universities (No. JZ2020HGQA0186), and the Open Foundation Project of Anhui Province Key Laboratory of Infrared and Cryogenic Plasma (No. IRKL2022KF06).

Data Availability Statement: Not applicable, the study does not report any data.

Acknowledgments: This work was partly supported by the Natural Science Foundation of China (No. 61806065 and No. 62120106008), the Fundamental Research Funds for the Central Universities (No. JZ2020HGQA0186), and the Open Foundation Project of Anhui Province Key Laboratory of Infrared and Cryogenic Plasma (No. IRKL2022KF06).

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

References

1. Wu, X.; Zhu, X.; Wu, G.Q.; Ding, W. Data mining with big data. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 97–107.
2. Shao, J.; Bu, C.; Ji, S.; Wu, X. A Weak Supervision Approach with Adversarial Training for Named Entity Recognition. In Proceedings of the Pacific Rim International Conference on Artificial Intelligence, Hanoi, Vietnam, 8–12 November 2021; pp. 17–30.
3. Bu, C.; Yu, X.; Hong, Y.; Jiang, T. Low-Quality Error Detection for Noisy Knowledge Graphs. *J. Database Manag.* **2021**, *32*, 48–64. [CrossRef]
4. Jiang, Y.; Wu, G.; Bu, C.; Hu, X. Chinese Entity Relation Extraction Based on Syntactic Features. In Proceedings of the 2018 IEEE International Conference on Big Knowledge, ICBK 2018, Singapore, 17–18 November 2018; pp. 99–105.
5. Li, J.; Bu, C.; Li, P.; Wu, X. A coarse-to-fine collective entity linking method for heterogeneous information networks. *Knowl.-Based Syst.* **2021**, *228*, 107286. [CrossRef]
6. Wu, X.; Jiang, T.; Zhu, Y.; Bu, C. Knowledge Graph for China's Genealogy. *IEEE Trans. Knowl. Data Eng.* **2021**, *1*, 1. [CrossRef]
7. Suchanek, F.M.; Kasneci, G.; Weikum, G. Yago: A large ontology from wikipedia and wordnet. *J. Web Semant.* **2008**, *6*, 203–217. [CrossRef]
8. Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: A collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, 10–12 June 2008; pp. 1247–1250.
9. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey, M.; Van Kleef, P.; Auer, S.; others. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semant. Web* **2015**, *6*, 167–195. [CrossRef]
10. Philpot, A.; Hovy, E.; Pantel, P. The omega ontology. In Proceedings of the OntoLex 2005-Ontologies and Lexical Resources, Jeju Island, Korea, 15 October 2005.
11. Xu, B.; Xu, Y.; Liang, J.; Xie, C.; Liang, B.; Cui, W.; Xiao, Y. CN-DBpedia: A never-ending Chinese knowledge extraction system. In Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Arras, France, 27–30 June 2017; pp. 428–438.
12. Wang, Z.; Li, J.; Wang, Z.; Li, S.; Li, M.; Zhang, D.; Shi, Y.; Liu, Y.; Zhang, P.; Tang, J. XLore: A Large-scale English-Chinese Bilingual Knowledge Graph. In Proceedings of the International semantic web conference (Posters & Demos), Sydney, Australia, 23 October 2013; Volume 1035, pp. 121–124.

13. Jiang, T.; Bu, C.; Zhu, Y.; Wu, X. Combining embedding-based and symbol-based methods for entity alignment. *Pattern Recognit.* **2021**, *2021*, 108433. [CrossRef]
14. Jiang, T.; Bu, C.; Zhu, Y.; Wu, X. Two-Stage Entity Alignment: Combining Hybrid Knowledge Graph Embedding with Similarity-Based Relation Alignment. In *PRICAI 2019: Trends in Artificial Intelligence—16th Pacific Rim International Conference on Artificial Intelligence, Cuvu, Yanuca Island, Fiji, 26–30 August 2019, Proceedings, Part I*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2019; Volume 11670, pp. 162–175.
15. Yan, Z.; Guoliang, L.; Jianhua, F. A survey on entity alignment of knowledge base. *J. Comput. Res. Dev.* **2016**, *53*, 165.
16. Newcombe, H.B.; Kennedy, J.M.; Axford, S.; James, A.P. Automatic linkage of vital records. *Science* **1959**, *130*, 954–959. [CrossRef]
17. Fellegi, I.P.; Sunter, A.B. A theory for record linkage. *J. Am. Stat. Assoc.* **1969**, *64*, 1183–1210. [CrossRef]
18. Herzog, T.N.; Scheuren, F.J.; Winkler, W.E. *Data Quality and Record Linkage Techniques*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2007.
19. Dong, X.; Halevy, A.; Madhavan, J. Reference reconciliation in complex information spaces. In Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, Baltimore, MD, USA, 14–16 June 2005; pp. 85–96.
20. Bhattacharya, I.; Getoor, L. Collective entity resolution in relational data. *Acm Trans. Knowl. Discov. Data* **2007**, *1*, 5. [CrossRef]
21. Maratea, A.; Petrosino, A.; Manzo, M. Extended Graph Backbone for Motif Analysis. In Proceedings of the 18th International Conference on Computer Systems and Technologies, Ruse, Bulgaria, 23–24 June 2017; pp. 36–43.
22. Pasula, H.; Marthi, B.; Milch, B.; Russell, S.J.; Shpitser, I. Identity uncertainty and citation matching. Presented at the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–13 December 2003; pp. 1425–1432. Available online: http://people.csail.mit.edu/milch/papers/nipsnewer.pdf (accessed on 17 December 2021).
23. Tang, J.; Li, J.; Liang, B.; Huang, X.; Li, Y.; Wang, K. Using Bayesian decision for ontology mapping. *J. Web Semant.* **2006**, *4*, 243–262. [CrossRef]
24. Bhattacharya, I.; Getoor, L. A latent dirichlet model for unsupervised entity resolution. In Proceedings of the 2006 SIAM International Conference on Data Mining, SIAM, Bethesda, MD, USA, 20–22 April 2006; pp. 47–58.
25. Hall, R.; Sutton, C.; McCallum, A. Unsupervised deduplication using cross-field dependencies. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV, USA, 24–27 August 2008; pp. 310–317.
26. McCallum, A.; Wellner, B. Conditional models of identity uncertainty with application to noun coreference. *Adv. Neural Inf. Process. Syst.* **2004**, *17*, 905–912.
27. Domingos, P. Multi-relational record linkage. In Proceedings of the KDD-2004 Workshop on Multi-Relational Data Mining, Citeseer, Washington, DC, USA, 22 August 2004.
28. Singla, P.; Domingos, P. Entity resolution with markov logic. In Proceedings of the Sixth International Conference on Data Mining (ICDM'06), Hong Kong, China, 18–22 December 2006; pp. 572–582.
29. Rastogi, V.; Dalvi, N.; Garofalakis, M. Large-scale collective entity matching. *arXiv* **2011**, arXiv:1103.2410.
30. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
31. Stoilos, G.; Venetis, T.; Stamou, G. A fuzzy extension to the OWL 2 RL ontology language. *Comput. J.* **2015**, *58*, 2956–2971. [CrossRef]
32. Sleeman, J.; Finin, T. Computing foaf co-reference relations with rules and machine learning. In Proceedings of the Third International Workshop on Social Data on the Web, Tokyo, Japan, 29 November 2010.
33. Zheng, Z.; Si, X.; Li, F.; Chang, E.Y.; Zhu, X. Entity disambiguation with freebase. In Proceedings of the 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, Macau, China, 4–7 December 2012; Volume 1, pp. 82–89.
34. Navarro, G. A guided tour to approximate string matching. *Acm Comput. Surv.* **2001**, *33*, 31–88. [CrossRef]
35. Smith, T.F.; Waterman, M.S. Identification of common molecular subsequences. *J. Mol. Biol.* **1981**, *147*, 195–197. [CrossRef]
36. Waterman, M.S.; Smith, T.F.; Beyer, W.A. Some biological sequence metrics. *Adv. Math.* **1976**, *20*, 367–387. [CrossRef]
37. Winkler, W.E.; Thibaudeau, Y. *An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 US Decennial Census*; Citeseer: Washington, DC, USA, 1991.
38. Winkler, W.E. *Overview of Record Linkage and Current Research Directions*; Bureau of the Census, Citeseer: Washington, DC, USA, 2006.
39. Raimond, Y.; Sutton, C.; Sandler, M.B. Automatic Interlinking of Music Datasets on the Semantic Web. In Proceedings of the Automatic Interlinking of Music Datasets on the Semantic Web, LDOW, Beijing, China, 28 April 2008.
40. Xiaohui, Z.; Haihua, J.; Ruihua, D. Property Weight Based Co-reference Resolution for Linked Data. *Comput. Sci.* **2013**, *40*, 40–43.
41. Junfu, H.; Tianrui, L.; Zhen, J.; Yunge, J.; Tao, Z. Entity alignment of Chinese heterogeneous encyclopedia knowledge base. *J. Comput. Appl.* **2016**, *36*, 1881–1886.
42. Weili, Z.; Yanlei, H.; Xiao, L. Instance Alignment Algorithm Between Encyclopedia Based on Semi-supervised Co-training. *Comput. Mod.* **2017**, *12*, 88–93.
43. Zhenpeng, L.; Mengjie, H.; Bin, Z.; Jing, D.; Jianmin, X. Entity alignment for encyclopedia knowledge base based on topic model. *Appl. Res. Comput.* **2019**, *11*, 1–8.
44. Yumin, L.; Dan, L.; Kai, Y.; Hongsen, Z. Weighted average Word2Vec entity alignment method. *Comput. Eng. Des.* **2019**, *7*, 1927–1933.

45. Jianhong, M.; Shuangyao, L.; Jun, Y. Multi-information Weighted Fusion Entity Alignment Algorithm. *Comput. Appl. Softw.* **2021**, *7*, 295–301.
46. Sun, M.; Zhu, H.; Xie, R.; Liu, Z. Iterative Entity Alignment Via Joint Knowledge Embeddings. In Proceedings of the International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017.
47. Pedrycz, W. *Granular Computing: Analysis and Design of Intelligent Systems*; CRC Press: Boca Raton, FL, USA, 2018.
48. Christen, P. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Trans. Knowl. Data Eng.* **2011**, *24*, 1537–1555. [CrossRef]
49. Griffiths, T. Gibbs Sampling in the Generative Model of Latent Dirichlet Allocation. 2002. Available online: https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.138.3760 (accessed on 17 December 2021).