International Journal of
Geo-Information

Article

Extracting Spatio-Temporal Information from Chinese Archaeological Site Text

Wenjing Yuan [1], Lin Yang [1,2,3,*], Qing Yang [1], Yehua Sheng [1,2] and Ziyang Wang [1]

[1] Key Laboratory of Virtual Geographical Environment, Ministry of Education, Nanjing Normal University, Nanjing 210023, China; 201302143@njnu.edu.cn (W.Y.); 191335001@njnu.edu.cn (Q.Y.); shengyehua@njnu.edu.cn (Y.S.); 191335025@njnu.edu.cn (Z.W.)
[2] Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing 210023, China
[3] Institute of Environment Archaeology, Nanjing Normal University, Nanjing 210023, China
* Correspondence: 09298@njnu.edu.cn; Tel.: +86-25-83598270

Abstract: Archaeological site text is the main carrier of archaeological data at present, which contains rich information. How to efficiently extract useful knowledge from the massive unstructured archaeological site texts is of great significance for the mining and reuse of archaeological information. According to the site information (such as name, location, cultural type, dynasty, etc.) recorded in the Chinese archaeological site text, this paper combines deep learning and natural language processing techniques to study the information extraction method for automatically obtaining the spatio-temporal information of sites. The initial construction of the corpus of Chinese archaeological site text is completed for the first time, and the corpus is input into the Bidirectional Long Short-Term Memory with Conditional Random Fields (BiLSTM-CRF) entity recognition model and Bidirectional Gated Recurrent Units with Dual Attention (BiGRU-Dual Attention) relationship extraction model for training. The F1 values of BiLSTM-CRF model and BiGRU-Dual Attention model on the test set reach 87.87% and 88.05%, respectively. The study demonstrates that the information extraction method proposed in this paper is feasible for the Chinese archaeological site texts, which promotes the establishment of knowledge graphs in archaeology and provides new methods and ideas for the development of information mining technology in archaeology.

Keywords: archaeological site; Chinese archaeological text; information extraction; deep learning

Citation: Yuan, W.; Yang, L.; Yang, Q.; Sheng, Y.; Wang, Z. Extracting Spatio-Temporal Information from Chinese Archaeological Site Text. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 175. https://doi.org/10.3390/ijgi11030175

Academic Editors: Susana Del Pozo, Jan Dirk Wegner, Lloyd A. Courtenay and Wolfgang Kainz

Received: 29 November 2021
Accepted: 2 March 2022
Published: 4 March 2022

1. Introduction

Archaeological sites are the remains of ancient human activities, which contain rich humanistic and social information and the law of civilization advancement. Archaeological site texts refer to the texts describing the information of archaeological sites, which is a significant carrier of site attribute information. Formally, it is often recorded discretely in various formats, for example, archaeological excavation reports, archaeological excavation briefings and archaeological dictionaries and encyclopedia entries in an unstructured manner. As far as quantity, with the consistent advancement in archaeological work, the textual data in archaeology are increasing, and more and more information on archaeological sites is being accumulated. In terms of content, the degree of detail varies in different types of archaeological site texts, but they all describe the basic information on the site (including name, location, dynasty, cultural type and other key elements), which is an important data hotspot for archaeological research and analysis. As a rule, the content of archaeological site texts primarily incorporates two perspectives: time and space. As the archaeologist Sqaulding said in his published book in 1960, "In short, archaeology is a science that concentrates on the form, time and spatial distribution of ancient remains" [1]. For archaeology, time and space are essential characteristics of coexisting with the form of remains [2]. The mutual contents contained in various archaeological site texts mentioned above are the descriptions of essential spatio-temporal information about the site. In archaeological texts, time information is a depiction of the site's historical period, which may be described as one or more dynasties or some cultural types to which the site belongs. Such descriptions of time are not uniform and might be precise or vague, so determining the period of the remains from the archaeological site text is a vital and fundamental assignment. Spatial information of the site's geographical location may be unequivocally described as an identified coordinate, an administrative region name or even a vague relative location. Therefore, it also expected to be recognized, interpreted and expressed uniformly in the information extraction. This paper concentrates on the extraction method of the sites' spatio-temporal information.

Archaeological site texts are the basis for archaeological site research, which contains rich information and research value. Therefore, in order to realize the effective utilization of archaeological site texts, it is especially critical to integrate and use the information on archaeological sites and mine the key and valuable knowledge of archaeology. The traditional method of manual identification to obtain site information from voluminous documents is time-consuming and inefficient, and the results of data structuralization may differ due to the inconsistent levels of various staff, which is inapplicable for information extraction from massive site texts. To date, the problem has received scant attention in the research literature, so there are few studies that have researched archaeological site texts. Therefore, how to extract unified site information from a large number of scattered, detailed or brief unstructured archaeological site texts is the crux of realizing the digitization and comprehensive utilization of archaeological site texts. In recent years, with the increasing development of artificial intelligence technology, information extraction methods and applications for natural language have made great progress. According to Cowie, Information Extraction can be defined as follows: 'Information Extraction (IE) is the name given to any process which selectively structures and combines data which is found, explicitly stated or implied, in one or more texts' [3]. The existing body of research on information extraction suggests that this technology is meaningful and promising. By analyzing the existing research on archaeological text information extraction, it is observed that these studies are basically oriented to English corpus, while Chinese studies are generally based on rules, which has poor portability and high implementation cost. The research on Chinese information extraction in archaeological site texts, which is limited by problems such as corpus annotation and Chinese word segmentation. At the same time, there are some deficiencies in these studies, such as singular data sources, fragmented construction processes and so on.

Under this foundation, this paper focuses on the extraction of spatio-temporal information from archaeological site texts. The information extraction experiment is mainly divided into two parts: entity recognition and relationship extraction. Their main goal is to recognize entities from texts and extract semantic relationships between entities. For a given input sentence, entity recognition involves both entity segmentation and the entity type. Relationship extraction aims to identify the semantic relations between symmetric entity pairs from unstructured archaeological site texts and to express them according to the structured form of a triplet (e1, r, e2), where e1 and e2 represent the first entity and the second entity, respectively, and r represents the relationship type between them. Finally, the temporal and spatial information of the site will also be presented in this form. In traditional natural language processing, entity recognition and relationship extraction are two independent tasks. The entity recognition model in this paper, named the Bidirectional Long Short-Term Memory with Conditional Random Fields (BiLSTM-CRF), combines the content of the application in natural language processing [4] and performs some research on data preprocessing and data analysis. Through the application of the BiLSTM-CRF model, it is able to effectively remember the context information and obtain the dependency relationship between adjacent tags, so as to obtain the optimal labeling results of an archaeological entity. In the relationship extraction task, the Bidirectional Gated Recurrent Units with Dual Attention (BiGRU-Dual Attention) model taken in this study is a mixed methodology based on previous studies [5,6]. For the task of Chinese relationship extraction, Chinese words, as the most basic unit in Chinese, contain a large amount of important semantic information. Therefore, the word-level information in Chinese training examples is very important for Chinese relationship extraction. With good optimization effect, the introduction of an attention mechanism can fully extract the context information of archaeological texts, so as to strengthen the extraction effect. The word-level attention mechanism and sentence-level attention mechanism in the model can better allocate weight, eliminate noise and improve the recognition accuracy of entity relationship extraction. By taking advantage of a neural network, the BiGRU-Dual Attention model can solve the problems of low accuracy and poor stability of traditional relationship extraction models. The main purpose is to reduce the work of manual processing and open up new directions and ideas of archaeological analysis.

In summary, the specific objective of this study was to rapidly and automatically identify and obtain the target information from a large amount of unstructured archaeological site texts by using new technology, thus greatly reducing the preprocessing time of archaeological information extraction. In addition, data for this study were collected from multiple carriers as comprehensively as possible, which provides new ideas and methods for the spatio-temporal information study of archaeological sites. The information extraction of archaeological site text makes an important contribution to the storage, management, utilization and sharing of archaeological knowledge and maximizes the value of archaeological site text.

2. Related Work

As referenced above, information extraction is the key technology of automatically extracting information from archaeological site texts. Around the early 1960s, the research of information extraction technology arose, and this technique empowers rapidly procuring target information from plentiful unstructured texts, bringing about a higher utilization of information. Information extraction methods generally include rule-based, statistics-based and deep-learning-based methods [7]. The exemplary LaSIE-II (Large Scale Information Extraction) system depends on semantic rules to realize information extraction [8]. However, this rule-based method has its own restrictions, such as the process of making rules manually being complex and the universality being poor. Consequently, the later research gradually turned to a statistics-based method. In a study conducted by Chambers, it was shown that the statistical learning algorithm can learn the rules from plain texts and perform the information extraction task without knowing the template structure in advance [9]. In the subsequent studies, researchers found that a method based on statistics is more viable than the previous method, but the cost of labor and time is extremely high since it additionally requires manual annotation with professional knowledge. Lately, the neural network models based on deep learning can automatically obtain feature information from a large number of texts, which provides direct support for the information extraction techniques. The model based on deep learning enormously outperforms the conventional methods in efficiency and accuracy and subsequently became applied broadly and gradually occupied the mainstream in information extraction tasks. Several studies of deep-learning-based information extraction have yielded fruitful outcomes. Qiu et al. proposed an Att-BiLSTM-CRF model based on an attention mechanism to effectively extract information entities in geoscience reports [10]. Zhang et al. implemented the structured course of geological entity information by utilizing a deep neural network [11]. Zhao combined the attention mechanism with the labeling and filtering layer in the Bidirectional Gated Recurrent Units (Bi-GRU) model, which significantly affects the relationship extraction of requirement text in the software industry [12]. From the current state of research, neural networks and CRF methods have become the de facto standard representing some of the best options for information extraction methods.

The application area of information extraction has gradually expanded with the development of its technology. The early research mainly focused on the study of textual information extraction tasks in general-purpose domains, such as the recognition of people's names and organizations' names [13,14]. On the basis of constant optimization in the general domain over the years, it has promoted the development of information extraction from texts towards more fields, which includes medicine, the military, agriculture and so on [15–17]. Simultaneously, information extraction has also developed towards a higher stage, such as relationship extraction, event extraction and other more complicated tasks [18,19]. Nowadays, it is observed that information extraction technology has also been explored and applied in history and humanities. For example, Sprugnoli proposed a neural method with manual annotation, which was applied in the place name recognition of English historical tourism texts [20]. What is more, Pettersson et al. put forward an online tool named HistSearch, which could effectively extract useful information from historical texts in a short time [21]. On the basis of the study on English archaeological reports at the previous stage, Vlachidis et al. developed the named entity recognition system of Dutch archaeological gray documents, which was able to achieve the semantic annotation of archaeological reports and automatically generate metadata [22]. With reference to the existing literature and codes, they are mostly for the English corpus and usually use word vectors for training.

The study of Chinese text information extraction mainly paid attention to named entity recognition at the initial stage. After that, it was gradually expanded to the tasks on the relationship and event extraction. In the interim, the field of information extraction was gradually expanded to a larger scope. In terms of the data mining of Chinese archaeological texts, it started relatively late, while it has also obtained some research results. For example, Zhang took advantage of the domain knowledge to carry out the extraction of data from archaeological texts [23]. However, it is difficult for this pattern-based method to learn enough text patterns, and would be mixed with a large number of meaningless word sequences. For the work adopting this method, it usually needs to be combined with complex verification and filtering. Lu made the proposal of a creative design platform for Changsha kiln cultural relics and extracted the text features of Changsha kiln cultural relic elements by using the BiLSTM-CRF model [24]. As a result, it achieved the construction of Changsha kiln cultural knowledge base. Based on deep learning technology, the platform realized the redesign of cultural relics elements, which promoted the integrated development of culture and technology. By combining Chinese word segmentation with entity recognition, Zhang effectively realized information extraction from archaeological text data [25]. However, he only carried out experiments on the data of Liangzhu site, which was lacking popularization and universality. Through the use of information extraction technology, Liu adopted the BiLSTM-CRF model to identify the entities such as person name, location name and time in the Twenty-Four Histories [26]. After that, he constructed the knowledge graph and stored the extracted knowledge through the neo4j graph database, which realized the semantic retrieval function. However, it still requires a lot of manual work involved in the classification of single and complex sentences when training dependent syntactic analysis models, which makes the model construction lack sufficient automation. Collectively, these studies indicate that information extraction technology based on deep learning has been studied in the field of Chinese archaeological site texts, but few studies have been able to draw on systematic research in the whole process. Meanwhile, such studies remain narrow in focus, dealing only with a specific object without generality. In addition, deep learning is the mainstream method at present, and its achievements have been remarkable.

To summarize, the study of information extraction has gone through decades from pattern recognition to machine learning to deep learning, from general field to professional domain, from regular standard text to ordinary text, and its achievements are remarkable. Based on the above analysis, the research on information extraction in Chinese Archaeology represented by named entity recognition and relationship extraction has made great progress, but it still has broad room for improvement in technology and methods. Firstly, compared with the general field, archaeological texts are rich in resources, but the information contained is complex. There are a large number of proprietary entities in the archaeological field, and it is difficult to identify them, so the research of information extraction focuses on its effectiveness and automation. In addition, archaeological texts put forward higher requirements for the accuracy of relationship extraction because of their complex syntax and the dense distribution of entity pairs with abundant overlapping relationships. Therefore, in view of the high complexity and domain specificity of Chinese archaeological texts, this paper uses natural language processing and deep learning methods to study entity recognition and relationship extraction in Chinese archaeological site texts. Moreover, it is hoped that this method can realize the processing of multi-source text data, complete the establishment from corpus to knowledge graph and truly complete the transformation from unstructured to structured data. According to the practical needs of archaeology, the named entity recognition is accomplished by the BiLSTM-CRF model, and the entity relationship extraction is completed by the BiGRU-Dual Attention model. Finally, the methods and techniques applicable to the archaeological site texts were experimentally tested, and the information extraction model for archaeological site text was constructed. The above study provides a new method for information acquisition in archaeology, which has important research value and application significance for promoting archaeological informatization.

3. Materials and Methods

3.1. Data

Chinese archaeology has an unrivaled assortment of valuable materials. Archaeological site texts are the primary vehicle for the presentation of results and academic exchange in archaeology, and its quantity has grown rapidly with the development of Chinese archaeological career. However, there is no publicly available corpus in the field of Chinese archaeology. Subsequently, taking the Chinese archaeological site texts as research data source, this paper gathers and organizes 625 Baidu Baike entries of sites [27], 300 archaeological excavation reports from CNKI [28], and 2325 entries from the Dictionary of Chinese Archaeology as the original data [29]. In the wake of arranging and summing up these data, we constructed a text corpus of Chinese archaeological sites. During the time spent researching, we observed that the text data in archaeology have their own characteristics compared with the text data in other fields. In terms of textual form, firstly, there will be some proper names that show up less often in other Chinese texts, such as '鬶' (a pitcher with three legs), '盉' (round vessel with a closed spout), '甗' (earthenware vessel) etc. Secondly, due to the various excavation methods of regional archaeological institutions, the workload and working conditions are different. Simultaneously, different archaeological recorders have different recording styles. According to these characteristics, we need to concentrate on appropriate information extraction methods to process them. From the textual content, although the content of the site texts varies among various data sources, they all contain fundamental information such as the name of the site, its location, dynasty and cultural type, which is the data basis for archaeological information extraction in this paper.

The temporal and spatial information is of great value for archaeological research. As far as temporal characteristics are concerned, each archaeological site has its own period, but the sites themselves (especially prehistoric sites) often lack clear time identification, so the year of most sites cannot be accurately determined. In the current archaeological chronology framework, the expressions of archaeological chronology generally include absolute age and relative age. According to the analysis on time in the text of Chinese archaeological sites, it tends to be observed that the chronological information is preferred to record in the way of relative age (such as Paleolithic age, Neolithic age and Western Zhou Dynasty, etc.). In addition, the site text likewise utilizes the archaeological culture (such as Yangshao culture, Hongshan culture, etc.) as the time stamp to record the age of the site. Archaeological culture refers to cultural sites belonging to the same era, distributed in the same area, and with a gathering of characteristic cultural relics and remains. In light of this, it has developed the basic space-time frame and a method for constructing historical narratives from archaeology. Consequently, this study extracts the cultural type in texts as the chronological information of sites. In terms of spatial feature, it is the identification of geographical location in site texts. From the depiction of spatial information, it can be divided into two categories of precise description and fuzzy description. In the precise description, the geographic coordinates of the site are recorded in the text, which can be directly extracted as the spatial information of the site. In the fuzzy description, it uses the natural language to describe the spatial location, predominantly including the names of administrative regions. This kind of spatial information has an obvious administrative hierarchy and subordinate relationship and is usually accurate to the village. Considering the above analysis of the Chinese archaeological texts, this study determines the temporal information of the site by integrating the relative age and cultural type in texts. Meanwhile, the administrative place name is extracted as the spatial information of the archeological site. The specific method is as follows.

3.2. Methodology

Methodologically, the information extraction technology is adopted to extract specific information from massive archaeological text data. The unstructured texts are processed and transformed into structured information. With respect to the information extraction model for archaeological site text in this study, it chiefly covers the BiLSTM-CRF, named the entity recognition model, and the BiGRU-Dual Attention relationship extraction model. The training of the named entity recognition model requires a large amount of annotation data. Since the experimental data cannot use the public annotation database on the Internet, the annotation of archeological site text is completed with YEDDA [30]. After data cleaning assignments such as removing exceptional symbols and futile URLs and retaining important punctuation marks, the text data are annotated manually. According to the above analysis of the archaeological texts, we first defined the archaeological entity. We pick the words or phrases with descriptive significance about the site, such as site name, cultural type, geographical location and historical dynasty, in the text as the archaeological entity, since they are all contents with specific meaning in the archaeological field. The BIO strategy is used to annotate the data. In the labeling process, character is the minimum labeling unit. BIO represents the category and position of archaeological entity, B addresses the head of the entity, I represents the middle position of the entity except the head, O addresses that this character does not belong to any entity category and X refers to the entity category. According to this strategy, each character can be marked as "B-X", "I-X" or "O". The relevant tags of the four categories of entities in the archaeological site text are shown in Table 1.

Table 1. Archaeological entity tag set.

Entity Category	Head Tag	Middle Tag
Site name	B-Site name	I-Site name
Cultural type	B-Cultural type	I-Cultural type
Geographical location	B-Geographical location	I-Geographical location
Historical dynasty	B-Historical dynasty	I-Historical dynasty

As per the relationship between the above four archaeological entities, we have defined four archaeological relationships, namely: Culture of the site, Location of the site, Dynasty of the site and None. In the process of carrying out information extraction experiments, it is primarily divided into two parts. Initially, the archaeological text is input into the named entity recognition model sentence by sentence. The trained model can identify archaeological entities of preset categories and output sentences containing entities. Then, the above outcome is input into the relationship extraction model, which finally obtains the entity relationship triplet (e1, r, e2), also known as "SPO triplet (subject, predicate, object)". For instance, the demonstration of information extraction on archaeological site text is shown in Figure 1.

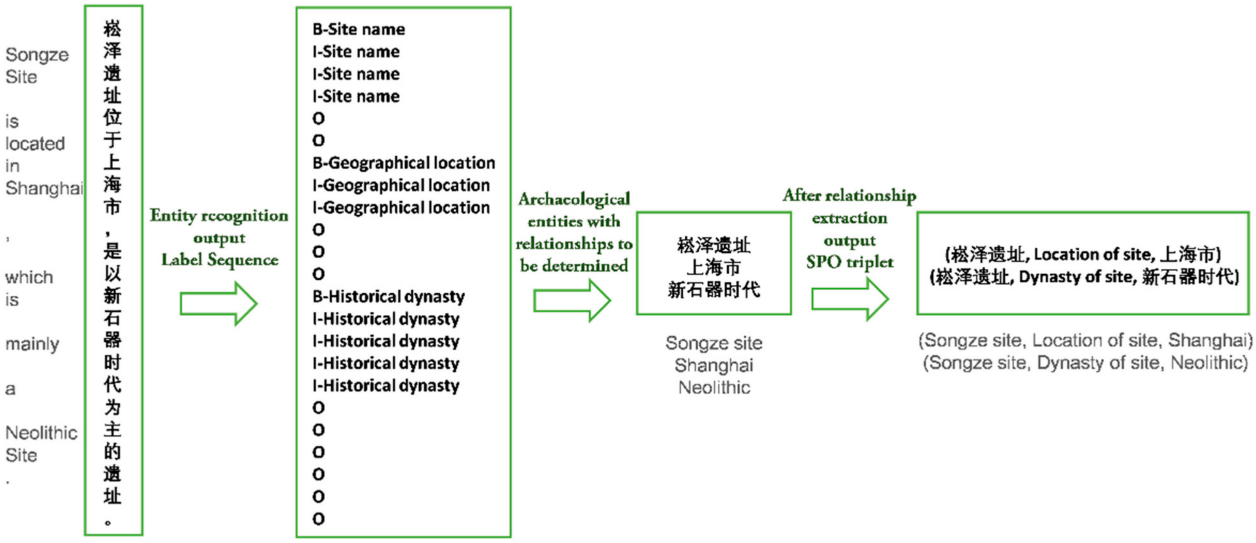
Songze Site is located in Shanghai , which is mainly a Neolithic Site .
崧泽遗址位于上海市，是以新石器时代为主的遗址。
Entity recognition output Label Sequence
B-Site name
I-Site name
I-Site name
I-Site name
O
O
B-Geographical location
I-Geographical location
I-Geographical location
O
O
O
B-Historical dynasty
I-Historical dynasty
I-Historical dynasty
I-Historical dynasty
I-Historical dynasty
O
O
O
O
O
O
Archaeological entities with relationships to be determined
崧泽遗址
上海市
新石器时代
Songze site
Shanghai
Neolithic
After relationship extraction output SPO triplet
(崧泽遗址, Location of site, 上海市)
(崧泽遗址, Dynasty of site, 新石器时代)
(Songze site, Location of site, Shanghai)
(Songze site, Dynasty of site, Neolithic)

Figure 1. Demonstration of information extraction on archaeological site text.

Named entity recognition is an essential assignment of extracting information. It refers to recognizing target entities in the text and classifying them as indicated by pre-defined criteria. In the experiment, the BiLSTM-CRF model is used for named entity recognition in Chinese archaeological site texts, which provides a novel thought and method for entity recognition in the field of Chinese archaeology. Entity relationship extraction means that after recognizing the vital entities in a sentence, the semantic relationships that exist between the entities are determined. Consequently, in light of entity recognition, this paper constructs a BiGRU-Dual Attention model for archaeological site texts. This model uses BiGRU to learn the contextual information of words to obtain finer-grained features. Through the word-level attention mechanism can increase the weight of words that are conclusive for relation classification. Simultaneously, using the sentence-level attention mechanism, we can learn more features of sentences and diminish the weight of noisy sentences, thereby effectively solving the problem of mislabeling and improving the effect of the classifier. The overall method process of this paper is shown in Figure 2.

3.2.1. BiLSTM-CRF Named Entity Recognition Model

LSTM is a kind of Recurrent Neural Network (RNN) for modeling text time series data. BiLSTM is a bidirectional LSTM, composed of a forward LSTM and a backward LSTM. However, BiLSTM can only predict the relationship between the text sequence and the tag and cannot predict the relationship between tags, so it requires the transition matrix in the CRF. In opposition to LSTM, the CRF can model hidden states and learn the characteristics of state sequences, but it needs to manually extract sequence features. Thusly, the BiLSTM-CRF model is constructed to obtain the upsides of both referenced previously.

The BiLSTM-CRF model constructed for recognizing archaeological named entities include four layers: input layer, embedding layer, BiLSTM layer and CRF layer. The specific structure of the entity recognition model is shown in Figure 3.

The first layer is the input layer, which takes the Chinese archaeological site text in words as the initial input, and a sentence containing n words is noted as $W = (w_1, w_2, w_3, \ldots, w_n)$, comprising a dictionary, where w_i is the id of the i-th word of the sentence in the dictionary, and the dimension is the dictionary size, which is the number of words.

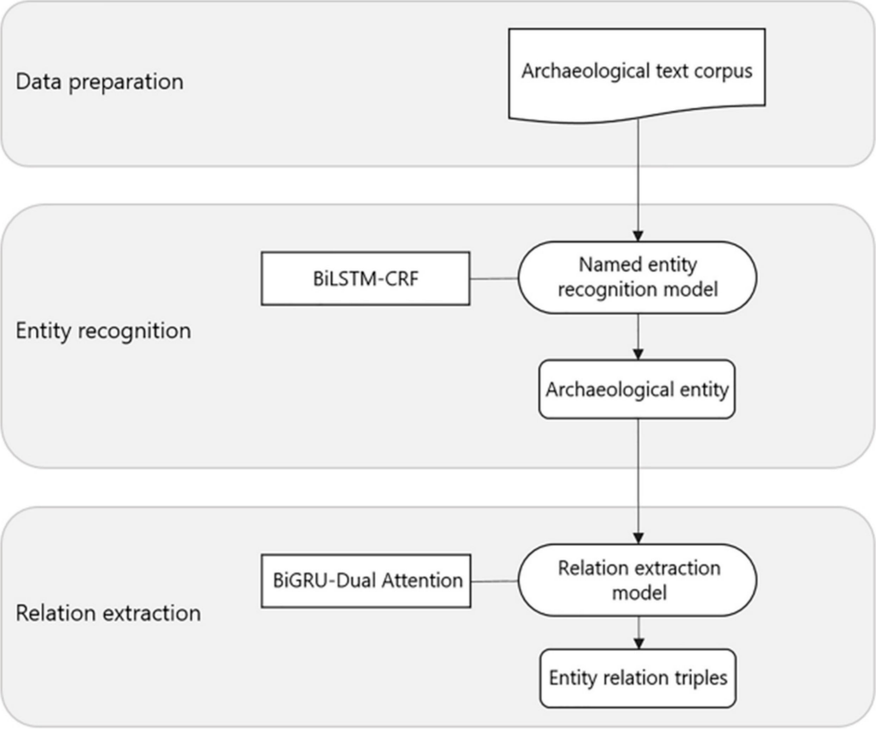

Figure 2. Flow chart of information extraction.

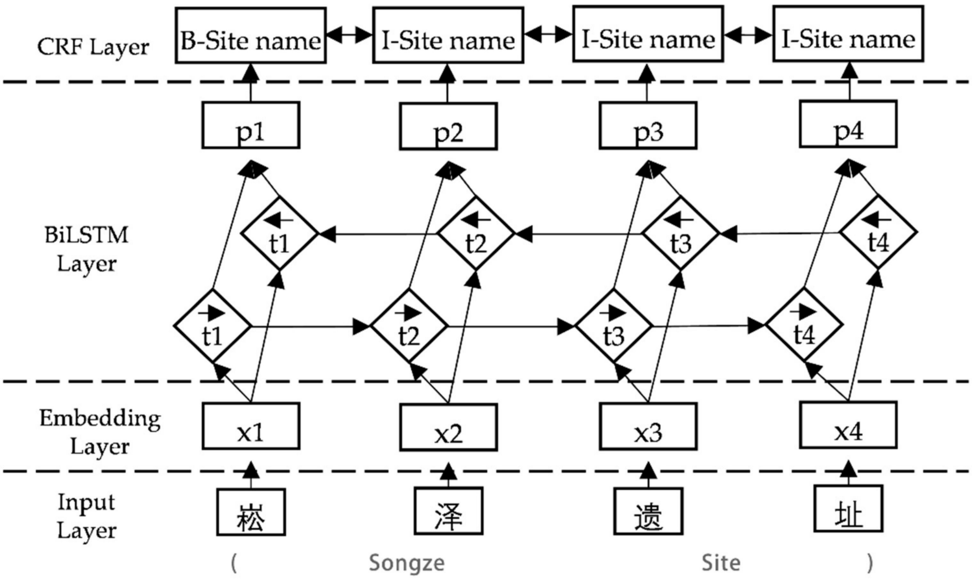

Figure 3. Bidirectional Long Short-Term Memory with Conditional Random Fields (BiLSTM-CRF) model.

The second layer is the embedding layer, which realizes the conversion of text data to computer-processable vector matrices through the word2vec tool. Each word is mapped into a word vector utilizing a random initialized matrix on this layer. For a given text sequence of unstructured archaeological sites, the word vector $X = (x_1, x_2, x_3, \ldots, x_n)$ is obtained.

On the BiLSTM layer, which consists of two LSTM layers, forward and backward semantic features are extracted according to the word vector input in each time step. Due to the difference in the sequence order of the vector processing, the two LSTM layers are divided into forward layer in positive order and backward layer in reverse order. The forward hidden layer is responsible for extracting the characterization of each word in the text and obtaining the output hidden state $T_1 = (\overrightarrow{t_1}, \overrightarrow{t_2}, \overrightarrow{t_3}, \ldots, \overrightarrow{t_n})$ of each word. The backward hidden layer is responsible for the reverse feature extraction, and the output hidden state $T_2 = (\overleftarrow{t_1}, \overleftarrow{t_2}, \overleftarrow{t_3}, \ldots, \overleftarrow{t_n})$ is obtained. Figure 3 shows the forward and backward propagation process and path through the arrow direction. At the same time, the BiLSTM network outputs the prediction scores of tags to the CRF layer, that is, $P = (p_1, p_2, p_3, \ldots p_n)$. Each dimension p_i of $p_{i,j}$ can be regarded as the score of classifying the word w_i into the j-th label.

The fourth layer is CRF layer, which considers the relationship between front and back words to control the annotation output order. Assuming that the input sentence W obtains a prediction tag sequence $y = (y_1, y_2, \ldots, y_n)$, the score of the prediction is defined as:

$$s(W, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{1}$$

where P_{i,y_i} is the probability that the BiLSTM output of the i-th position is y_i, and $A_{y_i,y_{i+1}}$ is the transition probability from y_i to y_{i+1}. The score of the whole sequence is the sum of the scores of each position. The score of each position is jointly determined by p_i and the transition matrix A of CRF. The score $s(W, y)$ of all possible annotation sequences y of W is obtained by the Viterbi algorithm, and afterward, all scores are normalized by the softmax function. Finally, the probability of sequence y is obtained as follows:

$$p\,(y|W) = \frac{e^{s(W,y)}}{\sum_{\overline{y} \in Y_W} e^{s(W,\overline{y})}} \tag{2}$$

While training the model, for the sentence input sequence X, the loss function is set to take the logarithm of the probability of the target real marking sequence Y. In order to maximize the probability corresponding to the real marker sequence, the strategy of taking a negative value and then minimizing it is taken on, and the gradient descent algorithm is introduced to solve the parameters. The maximize log likelihood function is as follows:

$$log(p\,(Y|X)\,) = s(X, Y) - log(\sum_{\overline{Y} \in Y_X} e^{s(X,\overline{Y})}) = s\,(X|Y) - log(\sum_{\overline{Y} \in Y_X} s\,(X|\overline{Y})\,) \tag{3}$$

In the prediction process, the S scores relating to all possible y sequences are calculated by the trained parameters, and the Viterbi algorithm is used to solve the optimal path. The predicted result is recorded as Y^*:

$$Y^* = \underset{\overline{Y} \in Y_X}{arg} \max(s(X, \overline{Y})) \tag{4}$$

3.2.2. BiGRU-Dual Attention Relationship Extraction Model

GRU is a variant of LSTM, which is simplified based on LSTM. Since the unidirectional GRU ignores the association between texts, BIGRU is used to carry out these associations in this study. In addition, this paper introduces the word-level and sentence-level dual attention mechanism, which can better eliminate noise interference and improve accuracy compared with the single-layer attention mechanism. The BiGRU-Dual Attention model is divided into six parts. The structure of the model is shown in Figure 4.

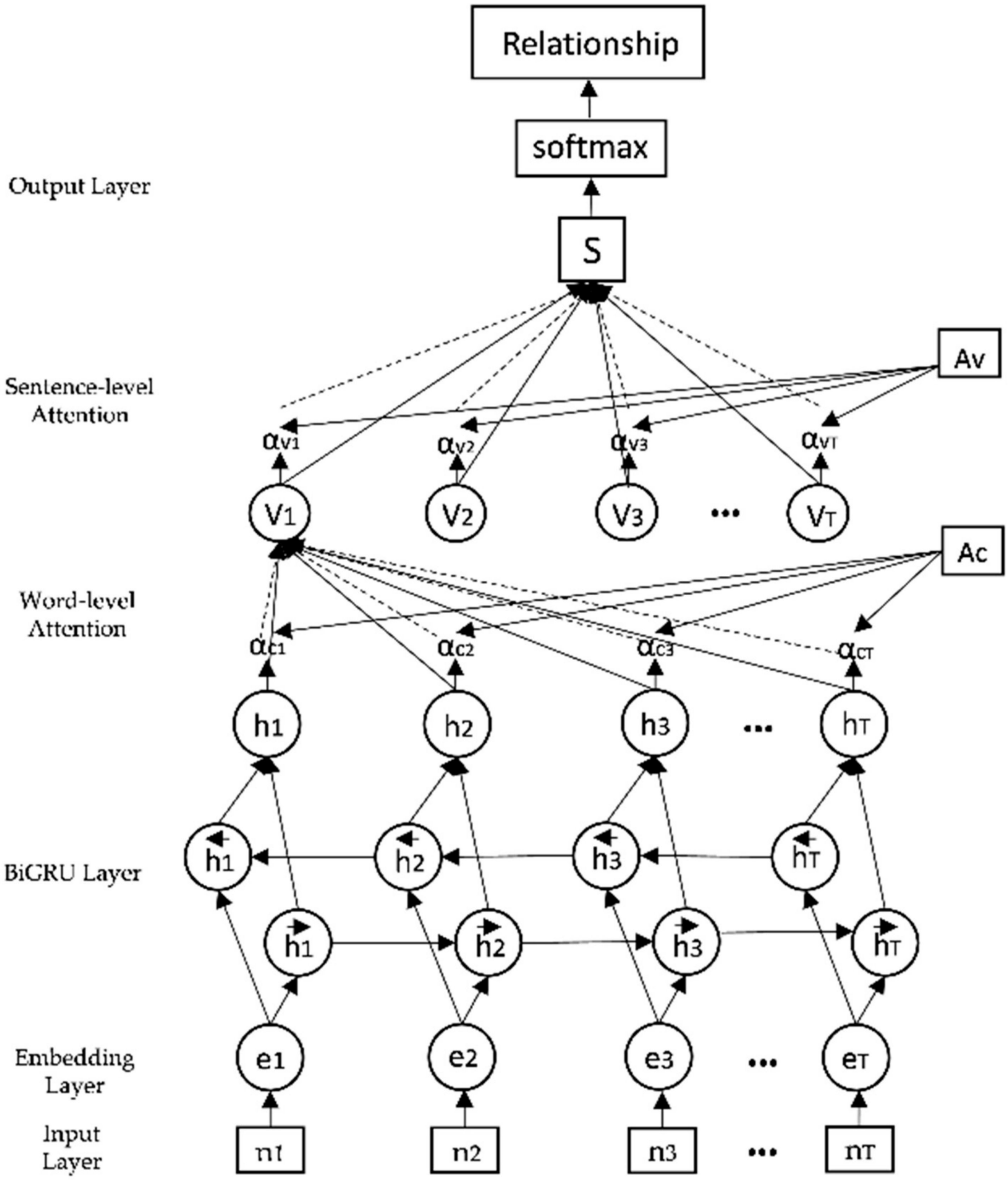

Figure 4. Bidirectional Gated Recurrent Units with Dual Attention (BiGRU-Dual Attention) model.

To start with, the input training instance $N = (n_1, n_2, n_3, \ldots, n_T)$ is transformed into the word vector sequence $E = (e_1, e_2, e_3, \ldots, e_T)$ through the embedding layer. Then, GRU is utilized to integrate the context information. Compared with the unidirectional GRU network, BiGRU adds one more hidden layer, which inputs the text sequence into the model in forward and reverse directions and connects the hidden layer states in both directions to the output layer. At this time, the network output corresponding to the *i*-th Chinese character is:

$$h_i = [\overrightarrow{h_i} \oplus \overleftarrow{h_i}] \quad (5)$$

where $\overrightarrow{h_i}$ is the output of the forward layer of GRU network with the word vector n_T as the input, $\overleftarrow{h_i}$ is the output of the reverse layer and $\oplus$ represents addition element by element.

For the word vector matrix $H = (h_1, h_2, h_3, \ldots, h_T)$ output by the BiGRU network, where T is the number of Chinese characters contained in the relationship instance.

Each word vector h_i is weighted by introducing the word-level attention weight A_C:

$$V = HA_C^T \tag{6}$$

where V is the calculated result vector, and A_C can be calculated by softmax function:

$$A_C = softmax\left(U_C^T M\right) \tag{7}$$

$$M = \tanh(H) \tag{8}$$

where U_C is the parameter used for training in the model, which is obtained in the training process.

The sentence-level attention mechanism takes the output of the word-attention mechanism layer as the input. By calculating the matching degree between each sentence containing entity pairs and the predicted relationship, the sentence level weight matrix is constructed, and finally, the vector representing the sentence is obtained. The specific algorithm flow is as follows:

$$S = \sum_i \alpha_{V_i} V_i \tag{9}$$

$$V_i = \tanh(V) \tag{10}$$

$$\alpha_{V_i} = \frac{\exp(k_i)}{\sum_j \exp\left(k_j\right)} \tag{11}$$

$$k_i = V_i A r \tag{12}$$

where S is the output vector of sentence-level attention mechanism layer, and α_{V_i} is the weight of each sentence vector V_i. The function k_i represents the matching degree between each sentence V_i and the predictive relationship r, and A is the weight diagonal matrix.

Then, the conditional probability $p(r|S)$ of the predictive relationship is calculated through the softmax function:

$$p(r|S) = softmax(RS + b) \tag{13}$$

where R is the matrix composed of all relation vectors, and b is the offset vector. Finally, the argmax function is used to obtain the relationship of the final prediction:

$$\hat{r} = argmaxp(r|S) \tag{14}$$

Based on Tensorflow, the paper realizes the relationship extraction model in light of dual attention mechanism, uses cross entropy as the loss function during training and combines L2 regularization to restrict the size of parameters to alleviate the problem of overfitting in the training process. The calculation of the loss function is as follows:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} t_i log(p(r_i|S_i, \theta)) + \lambda \|\theta\|_2^2 \tag{15}$$

where θ represents all parameters in the model, m represents the number of instance sets, t_i is the real relationship label and λ is L2 regularization coefficient. Then, the loss function is minimized by Adam algorithm to realize the training and stable updating of the parameters in the model. In addition, in order to avoid overfitting, dropout is added to the BiGRU layer.

4. Experimental Results

4.1. Experimental Setup

As referenced in Section 3.2, this study has identified four archaeological entities, including site name, cultural type, geographical location and historical dynasties. Next, the named entity recognition experiment is carried out with the above manually labeled corpus. A total of 21,800 corpora was selected from the texts of 800 sites as the experimental corpus of the named entity recognition experiment. Among them, 80% are used for the training corpus, 10% for the verification corpus and 10% for the test corpus. The statistics of entities in the datasets are shown in Table 2.

Table 2. Statistics of entities in the datasets.

Dataset	**Number of Entities**				
	Site Name	**Cultural Type**	**Geographical Location**	**Historical Dynasty**	**Total**
Training	3098	2100	6889	1766	13,853
Validation	405	236	769	210	1620
Test	416	210	828	191	1645

As far as the named entity recognition task of archaeological site text, experiments were conducted based on the Pytorch deep learning framework, and the parameter settings of the model in training are shown in Table 3.

Table 3. Training parameters of the BiLSTM-CRF model.

Parameter	**Value**
batch_size	64
learning rate	0.001
epoches	30
print_step	5
emb_size	128
hidden_size	128

On the basis of entity recognition, the relationship extraction of archaeological entities is carried out. For the entity relationship extraction in the field of archaeological texts, 8120 corpora are selected from the results of entity recognition, of which 80% are selected as training corpora, 10% as verification corpora and 10% as test corpora. The entity relationship involved are divided into four categories, including the Culture of the site, the Location of the site, the Dynasty of the site and None. The statistics of the relationships in the datasets are shown in Table 4.

Table 4. Statistics of relationships in the datasets.

Dataset	**Number of Relationships**				
	Culture of Site	**Location of Site**	**Dynasty of Site**	**None**	**Total**
Training	1768	3712	734	228	6442
Validation	216	478	92	29	815
Test	236	483	112	32	863

In this paper, the BiGRU-Dual Attention model for archaeological texts is constructed based on the Tensorflow deep learning framework, and the parameter settings of the model in training are shown in Table 5.

Evaluation is a necessary work in the fields of machine learning, natural language processing, information retrieval and so on, and the evaluation metrics are usually as follows: precision, recall and F1 value. Therefore, the information extraction model for archaeological site texts in this study uses the precision P, recall R and F1 value as the evaluation index. It can be calculated as follows:

$$precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{16}$$

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \tag{17}$$

$$\text{F1} = 2 * \frac{precision * recall}{precision + recall} \tag{18}$$

where true positives represent data that are truly predicted, false positives represent data that are incorrectly predicted and false negatives represent the data that should be correctly predicted but have not been predicted.

Table 5. Training parameters of the BiGRU-Dual Attention model.

Parameter	Value
vocab_size	166,919
num_steps	70
num_epochs	100
num_classes	4
gru_size	300
dropout	0.5
pos_size	5
pos_num	123
big_num	50

4.2. Entity Recognition Results

The BiLSTM-CRF entity recognition model was trained using the labeled archaeological site text. In order to evaluate the effectiveness of the BiLSTM-CRF entity recognition model in archaeological site texts, comparative experiments were conducted on a Hidden Markov Model (HMM, an early classical statistical model), a BiLSTM model with the same experimental data. The experimental results are shown in Table 6.

Table 6. Comparison of different entity recognition models in archaeological site text.

Model	P%	R%	F1%
HMM	87.12	75.96	81.16
BiLSTM	93.87	80.16	86.47
BiLSTM-CRF	94.51	82.10	87.87

From the analysis of the comparative experiment, the effect of the BiLSTM-CRF model constructed in this paper is superior to other methods, with a precision rate of 94.51%, a recall rate of 82.10% and an F1 value of 87.87%. This indicates that it has good adaptability in the entity recognition task of archaeological site text and can effectively carry out abstract modeling of archaeological texts. In terms of the three metrics, the model in this paper outperforms the HMM model, with an improvement in precision, in recall and in F1 value. It illustrates that the performance of the model relying on a neural network is obviously better than the early statistical model, with a significant improvement. The model in this paper improves the precision, the recall and the F1 value compared with BiLSTM, indicating that the addition of a CRF layer can effectively improve the recognition of relevant entities in the texts of archaeological sites.

In the following, further analysis about the recognition result of the BiLSTM-CRF model for various types of entities is shown in Table 7.

Table 7. Experimental results of BiLSTM-CRF model.

Type of Entity	P%	R%	F1%
Site name	95.31	70.86	81.29
Cultural type	98.64	90.38	94.33
Geographical location	94.52	86.80	90.50
Historical dynasty	88.17	77.12	82.28

It can be seen that this model can comparatively accurately recognize four types of entities in the archaeological site text, of which the F1 values of the geographical location and cultural type are above 90%. From the analysis of the experimental results, it can be observed that the precision of cultural type entities is the highest. This may be related to the clear identification of “culture” and “type” in Chinese archeological site texts, which is helpful to improve the recognition capacity of the model. Conversely, the description of historical dynasty is more complex in Chinese, so the model has difficulty finding a general rule expression, bringing about a relatively poor recognition result.

4.3. Relationship Extraction Results

In the experiment of entity relationship extraction, the precision P, recall R and F1 value are also used to evaluate the performance of the model. In order to verify the function of the BiGRU-Dual Attention model in precision and recall, the experimental result of relationship extraction from archaeological text entities is analyzed and compared with the BiLSTM-Attention model. The results are shown in Table 8.

Table 8. Comparison between BiLSTM-Attention and BiGRU-Dual Attention.

Model	P%	R%	F1%
BiLSTM-Attention	90.77	81.49	85.76
BiGRU-Dual Attention	91.83	84.64	88.05

The experimental results demonstrate that the BiGRU-Dual Attention model achieves better function than the BiLSTM-Attention model without increasing the complexity of the model. The BiGRU-Dual Attention model shows some improvement in performance, with progress in precision, in recall, and in F1 value. Meanwhile, it can be seen that the use of dual attention mechanism has a positive impact on improving the model performance and achieve higher precision in relationship extraction. In order to further analyze the difference in the extraction effect of various entity relationships, the evaluation results of different relationships are analyzed, as shown in Table 9.

Table 9. Experimental results of the BiGRU-Dual Attention model.

Type of Relationship	P%	R%	F1%
Culture of site	86.94	84.17	85.53
Location of site	93.98	86.37	90.01
Dynasty of site	93.79	81.54	87.24
None	88.47	72.90	79.93

Combined with the entity distribution of the labeled sample data set, it can be seen from Table 4 that the site location relationship accounts for the largest proportion in the test set, while the None relationship accounts for the least proportion. Relatively speaking, the relationship categories with a large amount of data has a higher recall rate during the test. From the above analysis, it can be seen that in the task of text relationship extraction, compared with the improvement of the model algorithm, the quality of the corpus is additionally vital. The higher the quality of the deep learning model’s training and learning sets, the more accurate the model recognition effect will be. In terms of effectiveness and feasibility, the comprehensive experimental results show that the BiGRU-Dual Attention model has a positive impact on relationship extraction in Chinese archaeological site texts.

The purpose of the above experiment is to demonstrate the feasibility of the application of information extraction in the field of archaeological site texts and aims to find a suitable method. Through the reflection on the test results, it can assist with enhancing the datasets and models in the follow-up research. Generally speaking, the BiLSTM-CRF model can effectively identify the four types of entities which relate to the spatio-temporal information of sites. However, it has low recall, which is caused by the changeability of sentence patterns in Chinese archaeological site texts. Later, we will add the entity-labeled corpus to improve the recognition ability of the model. On the basis of the entity recognition experiment, it is found that the BiGRU-Dual Attention model performs well in the task of archaeological site relationship extraction, which further enhances the training efficiency of the experiment. Furthermore, the reason for incorrect entity relationship recognition is mainly related to the lack of an annotation corpus, resulting in the lack of the relationship extraction ability of the model. In future research, the corpus of relational annotated texts will be expanded. We hope to improve the extraction ability of the model to provide a reference for constructing the knowledge graph of archaeological sites.

4.4. Application Example

Under the advancement of computers and the Internet over the years, it can be seen that knowledge graph technology has drawn extensive attention. Knowledge graphs have natural advantages for the analysis, display and utilization of the results of information extraction. As the structured semantic knowledge base, knowledge graphs can effectively process, handle and integrate massive amounts of information. Information extraction based on structured triples is an important step in the process of constructing knowledge graphs. After the above information extraction experiment, we obtain the triples from the archaeological site texts. By storing the triples in the relational database, we can obtain a basic knowledge graph and complete the transformation from unstructured texts to structured texts. In line with the knowledge graph construction process, the development and storage of archaeological site knowledge graphs are realized based on Neo4j. The graph contains 3318 nodes and 8120 edges in total. Figure 5 shows a partial knowledge graph of archaeological sites. Fundamentally, this paper aims to extract structured spatio-temporal information from various archaeological textual data and formalize them with a unified triplet representation. They support graphical query language access, so that deep knowledge can be obtained. The introduction of knowledge graphs is relatively new, and few studies have explored their application in the field of archaeology. In the future, we envisage linking archaeological site knowledge from different resources, and the utilization of these interrelated knowledge will further strengthen the discovery of archaeological knowledge. By constructing the archaeological site knowledge graph, it not only enriches archaeological site knowledge but also popularizes archaeology for the public. Meanwhile, it can establish the foundation for subsequent applications such as the semantic search for archaeological knowledge and intellectual question and answering.

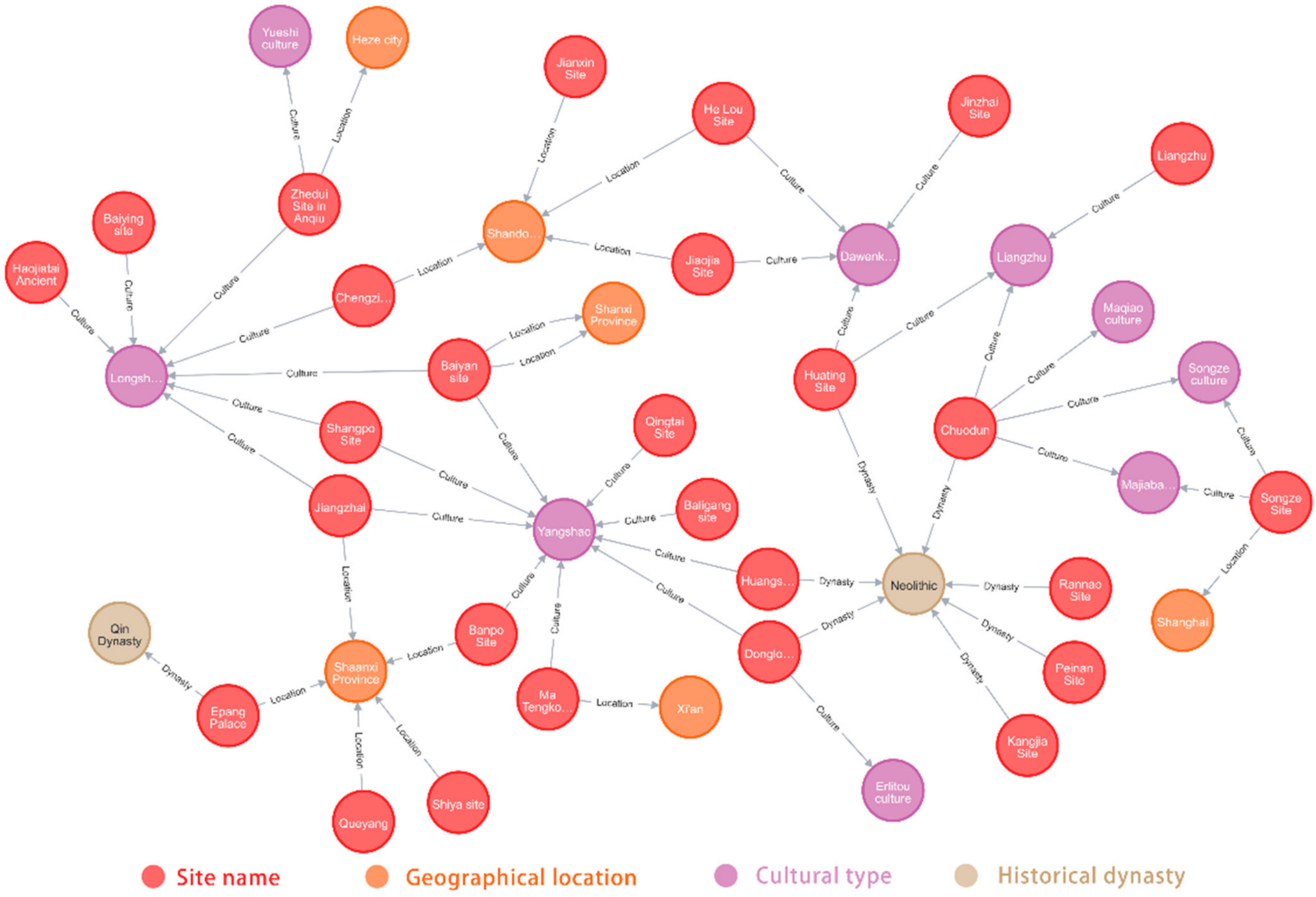

Figure 5. Instance of archaeological site knowledge graph (part).

5. Discussion and Conclusions

The archaeological site text is chosen as the research object in this study. Considering the issue of rich information with scattered knowledge in the field of Chinese archaeological site texts, its features and its application requirements are taken as the starting point. We utilize the information extraction method to extract the spatio-temporal information from the archaeological site text. The results show that it is suitable for relevant tasks. Compared with other existing studies, we explore the text of more data sources and naturally integrate them together. This study has obtained multi-source data, such as archaeological books, excavation reports and online texts. It has the benefit of obtaining higher-quality information and good coverage of archaeological site fields, which is vital for knowledge discovery and acquisition. We prove that information extraction technology is suitable for the field of Chinese archaeology, rather than only discussing a single text object. Compared with Zhang [25], under the same evaluation metrics, the P, R and F1 of our information extraction experiment are marginally lower, essentially in light of the fact that the input data in his study are semi-structured, while our input data are structured and come from various sources. The performance of the named entity recognition model is similar to that of Liu [26] but with higher precision and low recall. With the continuous emergence of new entities, to guarantee the quality of named entity recognition, we need to maintain dictionaries. When dictionaries are not detailed or domain rules are not complete, there are often the characteristics of high precision and low recall. Simultaneously, it is also observed that there is an unbalanced distribution of entity relations in the text of archaeological sites. In a text, there are often more descriptions of location and less descriptions of culture or dynasty. Therefore, in the case of model algorithm adaptation, effective data augmentation is expected to make the distribution of entity relations balanced, so as to improve the overall effect of information extraction. Nowadays, various information sharing media provide useful knowledge, so it is difficult to establish a final and complete knowledge base. However, different knowledge sources can complement each other. Compared with the method of description flow used by Zhang [23], triplets can connect knowledge from different sources and publish them in a unified way. At the same time, we introduced knowledge graphs and conducted a preliminary exploration. It allows users to make complex queries in the knowledge graph to promote knowledge connection and sharing. On this basis, it is able to provide data support for relevant scholars and provide new ideas for traditional information retrieval.

The study of the spatio-temporal information extraction method and its effectiveness verification of Chinese archaeological site texts is conducted in this paper. By fully using the multi-source and heterogeneous archaeological site text data on the Internet, this study conducts data annotation, which preliminarily completes the construction of the Chinese archaeological site corpus. Since there is no public annotation dataset or corpus in the field of Chinese archaeology, through the analysis of Chinese archaeological site texts, this study makes an appropriate definition of entity relationship hierarchy about site spatial-temporal information. Based on this, it establishes the data foundation for knowledge extraction of archaeological sites. By relying on the deep learning method that does not need manual feature extraction, the BiLSTM-CRF, named the entity recognition model, and the BiGRU-Dual Attention relationship extraction model are constructed to extract the spatial-temporal information on the site. After that, this study conducted comparative experiments, which obtained relatively good experimental results. These results show the possibility of applying this information extraction method to archaeological site texts. To further verify the extraction results of entity relationship triples of archaeological sites, an example of a knowledge graph was completed. Therefore, a new method is provided for the storage and display of traditional archaeological site knowledge. According to the results of the study, it can promote the relevant research of spatio-temporal information mining on sites and provide the basis for the construction of knowledge graphs in archaeology. Moreover, it is of great reference value for promoting the innovation of archaeological research methods and the exploration of archaeological problems in the information age. In the follow-up work, it is intended to annotate more archaeological site entities (including excavated artifacts, site area and so on) to expand the corpus, which tries to perfect the construction of knowledge graphs in the Chinese archaeological field and enrich the connotation of knowledge. In the meantime, we will continue to develop and research semantic search, intelligent Q&A and other upper-level applications based on the Chinese archaeological site knowledge graph.

Author Contributions: Wenjing Yuan: Conceptualization, methodology, validation, formal analysis, resources, data curation, writing—original draft preparation, writing—review and editing, visualization. Lin Yang: Conceptualization, validation, formal analysis, writing—original draft preparation, writing—review and editing, supervision, funding acquisition. Qing Yang: Methodology, validation, resources, data curation, writing—original draft preparation, writing—review and editing. Yehua Sheng: Methodology, formal analysis, writing—review and editing, supervision, funding acquisition. Ziyang Wang: Writing—review and editing, resources, supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant Nos. 41631175, 41471102) and the Major Research Project of the China National Social Science Fund (Grant No. 15ZDB054).

Informed Consent Statement: Not applicable.

Data Availability Statement: Some data presented in this study are available on request from the corresponding author. The data are not publicly available due to the confidentiality clause of some projects.

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

References

1. Spaulding, A.C.J.S. *Anthropological Papers*; Numbers 57–62. Bulletin 173; Bureau of American Ethnology, Smithsonian Institution: Washington, DC, USA, 1960; Volume 132, p. 888.
2. Zhang, G. *Kaoguxue Zhuanti Liujiang [Six Specialist Archaeology Lectures]*; Wenwu Chubanshe: Beijing, China, 1986.
3. Cowie, J.; Lehnert, W. Information extraction. *Commun. ACM* **1996**, *39*, 80–91. [CrossRef]
4. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint* **2015**, arXiv:1508.01991.
5. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-based bidirectional long short-term memory networks for relation classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; pp. 207–212.
6. Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; Sun, M. Neural relation extraction with selective attention over instances. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Long Papers), Berlin, Germany, 7–12 August 2016; Volume 1, pp. 2124–2133.
7. Guo, X.; He, T. Survey about Research on Information Extraction. *Comput. Sci.* **2015**, *42*, 14–17.
8. Humphreys, K.; Gaizauskas, R.; Azzam, S.; Huyck, C.; Mitchell, B.; Cunningham, H.; Wilks, Y. Description of the LaSIE-II system as used for MUC-7. In Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax, Virginia, 29 April–1 May 1998.
9. Chambers, N.; Jurafsky, D. Template-based information extraction without the templates. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 976–986.
10. Qiu, Q.; Xie, Z.; Wu, L.; Tao, L.; Li, W. BiLSTM-CRF for geological named entity recognition from the geoscience literature. *Earth Sci. Inform.* **2019**, *12*, 565–579. [CrossRef]
11. Zhang, X.; Ye, P.; Wang, S.; Du, M. Geological entity recognition method based on Deep Belief Networks. *Acta Petrol. Sin.* **2018**, *34*, 343–351.
12. Zhao, J. Research on the Application of Vocabulary Relation Extraction Method of Demand Entity Based on Bi-GRU. *J. Phys. Conf. Ser.* **2021**, *1748*, 032032. [CrossRef]
13. Zhao, J.; Wang, X.; Guan, Y. Comparing feature combination with features fusion in Chinese named entity recognition. *J. Comput. Appl.* **2005**, *25*, 2647–2649.
14. Ling, Y.; Yang, J.; He, L. Chinese organization name recognition based on multiple features. In Proceedings of the Pacific-Asia Workshop on Intelligence and Security Informatics, Kuala Lumpur, Malaysia, 29 May 2012; pp. 136–144.
15. Yang, Z.; Huang, Y.; Jiang, Y.; Sun, Y.; Zhang, Y.J.; Luo, P. Clinical Assistant Diagnosis for Electronic Medical Record Based on Convolutional Neural Network. *Sci. Rep.* **2018**, *8*, 6329. [CrossRef] [PubMed]
16. Xing, M.; Yang, C.-H.; Jin, L.-Y.; Bi, J.-Q. Research on the Construction and Application of Knowledge Graph in Military Domain. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Guangzhou, China, 20–21 September 2020; p. 012053.
17. Chen, Y.; Kuang, J.; Cheng, D.; Zheng, J.; Gao, M.; Zhou, A. AgriKG: An agricultural knowledge graph and its applications. In Proceedings of the International Conference on Database Systems for Advanced Applications, Chiang Mai, Thailand, 22–25 April 2019; pp. 533–537.
18. Leng, J.; Jiang, P. A deep learning approach for relationship extraction from interaction context in social manufacturing paradigm. *Knowl.-Based Syst.* **2016**, *100*, 188–199. [CrossRef]
19. Ritter, A.; Etzioni, O.; Clark, S. Open domain event extraction from twitter. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 1104–1112.
20. Sprugnoli, R. Arretium or Arezzo? A Neural Approach to the Identification of Place Names in Historical Texts. In Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-It 2018, Torino, Italy, 10–12 December 2018; pp. 360–365.
21. Pettersson, E.; Lindström, J.; Jacobsson, B.; Fiebranz, R. HistSearch-Implementation and Evaluation of a Web-based Tool for Automatic Information Extraction from Historical Text. In Proceedings of the HistoInformatics@ DH, Krakow, Poland, 11 July 2016; pp. 25–36.
22. Vlachidis, A.; Tudhope, D.; Wansleeben, M. Knowledge-Based Named Entity Recognition of Archaeological Concepts in Dutch. In Proceedings of the Research Conference on Metadata and Semantics Research, Madrid, Spain, 2–4 December 2020; pp. 53–64.
23. Zhang, C. A Research on Methods of Knowledge Acquisition from Domain-Specific Texts and Their Application in Knowledge Acquisition from Archaeological Texts. Master's Thesis, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, 2005.
24. Lu, W. Applying Deep Learning in Creative Re-creation of Changsha Kiln Cultural Relics. In Proceedings of the International Conference on Human-Computer Interaction, Copenhagen, Denmark, 19–24 July 2020; pp. 558–568.
25. Zhang, Y. Research and Application of Information Extraction and Analysis of Archaeological Excavations. Master's Thesis, Zhejiang University, Hangzhou, China, 2018.
26. Liu, R. The Construction and Retrieval of Knowledge Graph for the Biographical History Books. Master's Thesis, North University of China, Taiyuan, China, 2020.
27. Baidu Baike. Available online: https://baike.baidu.com (accessed on 30 November 2021).
28. CNKI. Available online: https://www.cnki.net (accessed on 30 November 2021).
29. Wang, W. *Dictionary of Chinese Archaeology*; Shanghai Ci Shu Chu Ban She: Shanghai, China, 2014.
30. Yang, J.; Zhang, Y.; Li, L.; Li, X. YEDDA: A lightweight collaborative text span annotation tool. *arXiv* **2017**, arXiv:1711.03759.