International Journal of Knowledge
and Language Processing AKLP International ©2010 ISSN 2191-2734
Volume 13, Number 2, Dec **2022** pp. 48—**62**

Construction of the Great Wall Cultural System and Annotated Corpus

Yajun Liu
Institute of Scientific and Technical Information of China
No. 15 FuXing Road, Haidian District, 100038
Beijing, China
540309659@qq.com

ABSTRACT: *The books related to the Great Wall culture contain significant knowledge about the Great Wall culture and history. Such books are characterized by their complexity, diverse content, intricate sentence structures, and rich wealth of knowledge. Since most of the original text data lacks labels, supervised model training cannot be conducted. The direct use of general models for text analysis in Great Wall books yields relatively poor results. This paper proposes a method for knowledge mining in Great Wall culture books, creating an annotated corpus that includes 11 types of entities and 21 types of semantic relationships. The paper explores the process of digitizing resources, annotation techniques, and analyzes the textual characteristics and language structures of books related to the Great Wall culture. It provides insights into constructing a corpus for the Great Wall culture, offering support for entity recognition and relation extraction, thereby laying a solid foundation for future research.*
Keywords: the Great Wall culture, knowledge mining, deep learning

1 Introduction. The Great Wall is an ancient military defense structure in China. It is a tall, sturdy, and continuous barrier used to fend off enemy attacks. The Great Wall is not just an isolated wall and it is a defense system that incorporates walls as the primary structure along with numerous forts, barriers, pavilions, and markers. Constructed from the Western Zhou period to the Ming Dynasty, its geographical span covers mainly Hebei, Beijing, Tianjin, and 15 other provinces, autonomous regions, and municipalities. Consequently, descriptions of the Great Wall in books are diverse and complex due to the vast geographical range, resulting in differences in language and culture. Some books may involve dialectal content, intricate sentence structures, diverse formats, and extensive knowledge. To study the Great Wall culture, a broad knowledge scope and an understanding of the cultural background of the Great Wall region are required.

Moreover, there is a lack of annotated data in the field of Great Wall culture and no publicly available dataset suitable for evaluating methods as a test set. Therefore, resources in this domain need to be constructed independently. Identifying and annotating various relevant concepts from Great Wall books pose significant challenges. With the advent of the big data era, employing text analysis methods to automatically extract information about the Great Wall culture from vast semi-structured and unstructured book data becomes a practical solution for establishing and maintaining a knowledge service platform in the Great Wall culture

domain. However, knowledge dispersed across various books and digital resource platforms exhibits diverse manifestations, large quantities, and complex content. Relying solely on manual methods for processing is undoubtedly extremely difficult. Furthermore, there is no standardized method for defining Great Wall cultural knowledge, restricting the extraction of entities and relationships. Therefore, formulating a refined and high-quality data annotation system to integrate knowledge into a unified framework is the primary challenge in excavating knowledge about the Great Wall culture.

Specific domain information extraction typically requires defining an ontology framework [1] that aligns with downstream tasks. Its significance lies in defining the boundaries and types of information to be extracted, offering standardized guidance for manual annotation and corpus construction. However, there is scarce research available in the Great Wall culture domain regarding ontology construction for reference in annotation cases. This fundamental limitation can be attributed to two main reasons: Firstly, at the data acquisition level, a massive amount of text data forms the fundamental element for knowledge mining. Various encyclopedic websites employ crowd sourcing to gather text data, often exceeding billions, such as electronic medical records in the medical field, financial statements in finance, and legal documents in the judicial field, providing rich data resources for information extraction. In contrast, the level of information in the field of the Great Wall culture is relatively low, lacking the ability to retain a large amount of text data during production processes. Moreover, domain knowledge bases compiled by relevant experts are scattered across various knowledge service platforms, lacking a unified standard to integrate these isolated network resources. Secondly, in terms of corpus annotation, the knowledge structure within the texts related to the Great Wall culture is mostly unstructured or semi-structured. Their distinctive features include entities and relationship types that are challenging to define. Entities often include a mix of numbers, dialects, special characters, etc. Consequently, the descriptions of relationships are ambiguous, leading to potential ambiguities, which pose significant challenges for manual annotation.

1. Related Work. Entities form a crucial part of studying the Great Wall culture. Without entities, the knowledge system of the Great Wall culture cannot be constructed. At the same time, the entities in the Great Wall books do not exist in isolation, and cultural veins run through each entity to form a network of the entire Great Wall cultural map, so it is crucial to accurately identify the entities to analyse the Great Wall cultural books. Named Entity Recognition (NER) is one of the foundational tasks in information extraction. With the advancement of big data and deep learning, the application of pre-trained language models has significantly improved NER model performance. In 2018, the pre-trained language model BERT [3] emerged and was applied to the field of entity recognition [4]. It was found to have a more significant effect in the Chinese domain, quickly becoming the primary method used in NER.

In academic research, entity naming recognition is generally divided into three

major categories (entities, events, and numbers) and seven subcategories (names of people, places, organizations, times, dates, currencies, and percentages). In practical applications, such as in news texts, entity naming recognition requires identifying the names of people, places, organizations, time, etc. In certain specific professional fields, such as in the medical field, entity recognition has been widely applied to electronic medical records, which often requires the recognition of proper nouns in the field, such as the name of the disease and the drug, etc. [5].

Books related to the Great Wall belong to the literary domain. Due to the complexity and variability of entities in literary works, coupled with the lack of annotated data, previous research on Named Entity Recognition (NER) has mostly focused on general domain corpora. Currently, there have been relatively few attempts to recognize named entities in the literary domain. Vala et al. [6] proposed a graph-based pipeline model to identify characters in literary works, achieving good results on multiple datasets and offering an evaluation framework for character identification performance, providing evaluation standards for future work. Brooke et al. [7] proposed a LitNER model for novel named entity recognition, which is based on the Bootstrap method and trained using an unsupervised approach, and experiments show that the model outperforms supervised methods given the full context of the text. Xu et al. [8] addressed the scarcity of NER data in Chinese literary works by constructing an NER and relationship extraction dataset based on over 700 literary articles. Xie Tao [9] conducted research on ancient texts such as *Song Ci* and *Shiji*, using the Apriori algorithm and LSTM model for tokenization and LSTM and CRF models for NER, effectively identifying important entities in ancient texts. Bamman et al. [10] annotated 100 English novels according to ACE annotation standards, enhancing the effectiveness of NER in the literary field.

The performance of named entity recognition systems is affected by a variety of factors, including linguistic factors, text type factors, and entity types. For Chinese named entity recognition, language is generally regarded as a continuous sequence of single words, and the sequence annotation method is used for entity recognition. The domain of the text being used for NER significantly affects the performance of the entity recognition system. In general, NER in professional domains is lower than in general domains, such as in the medical field [11]. Entity types also have a significant impact on NER systems. Entity types in general domains, such as personal names and dates, are relatively easier to identify and have higher accuracy, whereas Entities in specialized fields, like drug names or chemical terms, rely heavily on domain-specific textual resources and are challenging to generalize. They often involve nested entity types, making accurate identification difficult.

The research in this paper focuses on the Great Wall culture, encompassing entities related to this field such as archaeological sites, folk customs, as well as common entities found in historical texts like dynasties and nations. These entities have rich semantics, with some being highly specialized, and many involve nested entity types. For instance, The name of the Great Wall site is "Inner Mongolia Baotou Zhao Great Wall Site", which contains entities such as regions and Great Walls of past dynasties, making it difficult to identify.

This article mainly focuses on the field of Great Wall culture. The entities involved in the Great Wall book text not only include the basic types in academic research, such as historical figures, historical events, Great Wall sites, periods, regions and other entity events, but also include entities in their fields. Unique professional nouns such as "Puppet Yangge", "Henpecked Lamp", "Fire Spoon" and other nouns unique to the culture in folk customs, therefore it is difficult to achieve good indexing results by relying only on basic entity recognition methods. The main reasons are: first, the corpus in literary books is quite different from the corpus in general fields. Different authors have different writing styles, making it difficult for the model to generalize. Second, the entities in the work are complex and diverse, making it difficult for the model to learn features. Third, there are few studies on named entity recognition in this field, and there is a lack of large-scale and high-quality annotation corpus.

2. Model Construction. This chapter focuses on conceptual indexing of books related to the Great Wall culture. It introduces prior knowledge through domain pre-trained language models to enhance the model's adaptability in specialized domain texts. It constructs a conceptual indexing model that integrates prior conceptual knowledge. This is achieved by fine-tuning the BERT pre-trained model on domain-specific databases and jointly training it with the prediction task of conceptual entities. The aim is to identify known concepts within unstructured text and complete their conceptual categorization.

2.1 Domain Text Language Pre-trained Model. The research employs 94 books on the Great Wall culture as its corpus, encompassing a range of comprehensive and distinctive works, such as landscape stories, historical culture, local records, heritage series, cultural relics records, dictionaries, and concise histories related to the Great Wall. These books are stored in XML format and possess standardized organizational structures within their chapters.

Based on the corpus of Great Wall culture books mentioned above, the research fine-tuned the BERT model pre-trained on modern Chinese text. Based on bert-base-chinese, a BERT model specifically adapted to the domain of Great Wall culture books was obtained. Training was performed using the Chinese pre-trained BERT model hosted in Hugging Face's model hub along with the provided MLM fine-tuning script. The model parameters were set in line with those released by Google, with a batch size of 3 and fine-tuning epochs set to 10. The lowest loss was achieved at 30,000 steps, taking approximately 3 hours. The resulting language model embodies the textual characteristics of the Great Wall domain, incorporating domain-specific prior knowledge to enhance its adaptability to Great Wall culture texts. This trained model is intended not for direct use but as a foundation for training other tasks.

2.2 Construction of the Great Wall Text Concept Model. The concepts found in Great Wall books have two main characteristics: Firstly, they are highly correlated with the structure of Great Wall cultural knowledge, containing interrelated structures. Secondly, these concepts possess attributes that mutually describe and interact with

each other.As specialized domain text, Great Wall books typically contain numerous domain-specific and strongly stylized terminologies, varying text lengths from single characters to over ten characters for concepts. For instance, within the category of Great Wall architecture, concepts like "空" (representing a defensive structure of the Great Wall) and multi-character concepts such as "The Qin Emperor Zhao's Great Wall Site at Baijialiang, Inner Mongolia." are found. Concepts are not just limited to words or named entities but also include nested concepts. Multi-character concepts mentioned earlier are often nested, posing difficulty in entity recognition for excessively short or long concepts.

The concepts in book texts are highly related to the authors' backgrounds, writing habits, and textual structures within each book. Therefore, segmenting and selecting representative texts from Great Wall books for analyzing and modeling the contained concepts, and determining their conceptual categories, plays a crucial role in serving as training material for the concept indexing model. This study selects representative Great Wall books from existing literature. The conceptual modeling is based on the "Great Wall Dictionary," identifying specific conceptual terms and their types within the selected individual texts. For example:

The Great Wall built by the Qi State during the Warring States period. Qi was one of the vassal states of the Zhou Dynasty in the 11th century BC, located in the northern part of Shandong Province. ... The Qi Great Wall was constructed by connecting the embankments of the Ji River with mountain ranges to form the wall. It was built in two segments from west to east, gradually over the years. Its course started from the ancient Ji River in the north of Pingyin County, Shandong Province.

example 0. A text excerpt extract from "Great Wall Dictionary"

As a comprehensive book, the 'Great Wall Dictionary' contains rich conceptual semantics and covers a wide range of content. The writing in the book is relatively structured." By analyzing the excerpt above, it is evident that the text initially introduces the construction dynasties of the Great Wall and subsequently provides an explanation of the regions covered by this Great Wall. Analyzing the concepts, the Great Wall can be described with related entity concepts such as historical periods, regions, people, buildings, etc. Some concepts are annotated by entities, such as "Qi Guo", "Shandong Province" and other entity concepts in the above example, and other concepts are annotated by trigger words, such as "Building from the west", "Building". The word is the trigger word of the event.

In 1550, the twenty-ninth year of Jiajing in the Ming Dynasty, the Gengxu year of the Ganzhi calendar, the leader of the Mongolian Tumote tribe, Ada Khan, launched a war to invade the Ming Dynasty due to the failure of the "tribute market", which was known in history as the Gengxu Rebellion.

example 1. A text excerpt extract from "A Brief History of the Great Wall"

In the same way, it can be seen from the above example that the description of the "Gengxu Revolution" includes events, characters, and event trigger words. "1550, the 29th year of Jiajing in the Ming Dynasty, the year of Gengxu," is the time entity, "Ada Khan, the leader of the Mongolian Tumote Tribe" is the character entity, and "launching a war to invade the Ming Dynasty" is the event trigger word.

As can be seen from the above examples, the analysis of concepts usually focuses on the description of the concept theme. There are many types of concepts involved in the Great Wall, so it is crucial to conduct structured analysis of each type of concept, extract highly characteristic corpus, and analyze its unique language structure and semantic information. Based on the above analysis, this study combined multiple books with ontology dictionaries, and finally constructed a conceptual model as shown in Table 1.

Concept Category	Description	Example
People	Relevant figures	Wei Huiwang, Qin Shihuang
Folk customs	Local customs within the Great Wall region	Bamboo horse, Walking on stilts
Period	Dynasties and Periods	Tang Dynasty, Song Dynasty
Region	Names of regions	Yanqing, Inner Mongolia
Sites	Sites related to the Great Wall	Yulin Ancient Great Wall Site in Shaanxi, Fushun Pass
Historical Great Wall	Great Walls built in different dynasties	Ming Great Wall, Chu Great Wall
Historical Documents	Books and historical records	Examination of the Great Wall, Book of Northern Qi
Events	Events related to the Great Wall	Yongjia Rebellion, forcing the Song Dynasty to cross the river
Construction	Buildings and components	rammed earth, willow border
Significance of the Great Wall	Evaluative language regarding the Great Wall	Playing an increasingly important role, reflecting the great achievements of ancient architectural technology
Official Positions	Official titles of historical figures	Honored for his outstanding achievements as Wuan Jun, appointed as General of Conquest
Irrelevant	Descriptive and meaningless	Exactly how, so there is still doubt

Table 1: Concept Model of Great Wall Culture Books

2.3 Model Construction Incorporating Prior Conceptual Knowledge. This study incorporates prior knowledge based on a priori conceptual structure into the model, and this conceptual knowledge incorporates its unique triplet structure. During the training process, the model not only learns the textual features of the concept

knowledge, but also learns through the triplet structure contained in the concept to integrate deeper semantic-level features for semantic analysis.

Based on the pre-trained language model constructed previously, this section builds a model that incorporates prior knowledge. For the training corpus, the model randomly replaces the entities in it with wrong entities, and provides the model with correct entity information based on MarginRank-ingLoss, as shown in formula (1).

$$loss(x1,x2) = max(0, -(x1 - x2) + margin) \#(1)$$

where x is the vector output by the model, and $x2$ is the vector of randomly replaced error entities. Based on MarginRank-ingLoss, the quantity predicted by the model is closer to the vector of the real entity than the vector of the wrong entity, thereby obtaining the information of the entity in the sentence. The article uses a small sample of named entity corpus annotated from Great Wall books for model training.

During the training process, the study does not use the NSP loss function, but combines the MarginRank-ingLoss used in this section with the CrossEntropyLoss of MASK character prediction for training. Thus, the prior conceptual knowledge contained in the books is integrated into the model in a multi-task learning manner. During the model training process, the study adopts the same parameter settings as in the previous section, and obtains the optimal model at step 2460.

3 Experimentation and Evaluation.

3.1 Experiment Preparation. When running the indexing model that incorporates prior conceptual knowledge, the domain text language pre-training model trained in the previous article is used to obtain a pre-trained language model that contains the Great Wall domain knowledge. This model introduces domain prior knowledge and combines the domain text language pre-training model with the domain prior knowledge. The model is applied to a concept indexing model that integrates prior conceptual knowledge, which can enhance the adaptability of the model to texts in the Great Wall domain. The training corpus of the domain text language model based on transfer learning is the content of Great Wall Culture books, as shown in the figure below. The final generated model is step 30000. The figure below shows the final generated model file.

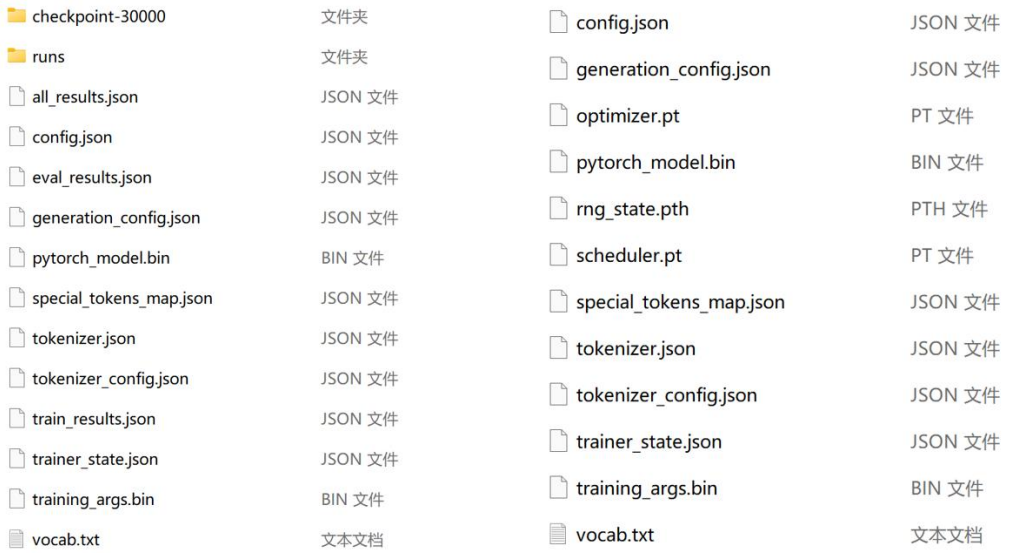


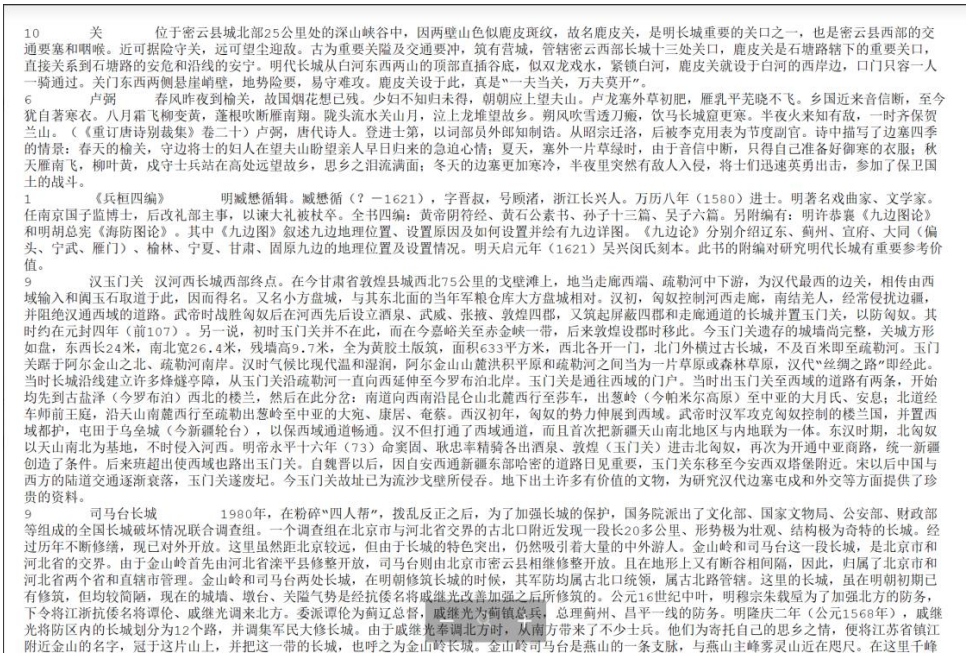
Figure 1:The Great Wall Text Pre-trained Language Model

The model, serving as the foundation for training the concept indexing model, needs to be incorporated during the training of the concept indexing model. Initially, a vast amount of domain knowledge is extracted from the Great Wall books to serve as the training corpus for the concept indexing model. This corpus does not contain relatively structured language patterns, and there are no constraints on the length of each text, formatted in .txt files, totaling 1583 text samples, as shown in the table below for each type of text sample.

Serial Number	Concept	Quantity
1	Historical Documents	138
2	people	201
3	Folk customs	106
4	Period	171
5	Region	256
6	Significance of the Great Wall	201
7	Historical Great Wall	79
8	Events	151
9	Sites	167
10	Constructions	113
11	Official Positions	150
12	literature	121
Total		1851

Table 2: The number of concepts in the training corpus.

Dividing the dataset, splitting it into a training set and a validation set in an 8:2 ratio, as illustrated in Figure 2, with the format of the corpus set as depicted.


10 关 位于密云县城北部25公里处的深山峡谷中，因两壁山色似鹿皮斑纹，故名鹿皮关，是明长城重要的关口之一，也是密云县西部的交通要塞和咽喉。近可据险守关，远可望生迎敌。古为重要关隘及交通要冲，筑有营城，管辖密云西部长城十三处关口，鹿皮关是石塘路辖下的重要关口，直接关系到石塘路的安危和沿线的安宁。明代长城从白河东西两山的顶部直插谷底，似双龙戏水，紧锁白河，鹿皮关就设于白河的西岸边，口门只容一人一骑通过。关门东西两侧悬崖峭壁，地势险要，易守难攻。鹿皮关设于此，真是"一夫当关，万夫莫开"。

6 卢弼 春风昨夜到榆关，故国烟花想已残。少妇不知归未得，朝朝应上望夫山。卢龙塞外草初肥，雁乳平芜晓不飞。乡国近来音信断，至今犹自著寒衣。八月霜飞柳变黄，蓬根吹断雁南翔。陇头流水关山月，泣上龙堆望故乡。朔风吹雪透刀瘢，饮马长城窟更寒。半夜火来知有敌，一时齐保贺兰山。（《重订唐诗别裁集》卷二十）卢弼，唐代诗人。登进士第，以词部员外郎知制诰。从昭宗迁洛，后被李克用表为节度副官。诗中描写了边塞四季的情景：春天的榆关，守边将士的妇人在望夫山盼望亲人早日归来的急迫心情；夏天，塞外一片草绿时，由于音信中断，只得自己准备好御寒的衣服；秋天雁南飞，柳叶黄，戍守士兵站在高处远望故乡，思乡之泪流满面；冬天的边塞更加寒冷，半夜里突然有敌人入侵，将士们迅速英勇出击，参加了保卫国土的战斗。

1 《兵桓四编》 明臧懋循辑。臧懋循（？－1621），字晋叔，号顾渚，浙江长兴人。万历八年（1580）进士。明著名戏曲家、文学家。任南京国子监博士，后改礼部主事，以谏大礼被杖卒。全书四编：黄帝阴符经、黄石公素书、孙子十三篇、吴子六篇。另附编有：明许恭襄《九边图论》和明胡总宪《海防图论》。其中《九边图》叙述九边地理位置、设置原因及如何设置并绘有九边详图。《九边论》分别介绍辽东、蓟州、宣府、大同（偏头、宁武、雁门）、榆林、宁夏、甘肃、固原九边的地理位置及设置情况。明天启元年（1621）吴兴闵氏刻本。此书的附编对研究明代长城有重要参考价值。

9 汉玉门关 汉河西长城西部终点。在今甘肃省敦煌县城西北75公里的戈壁滩上，地当走廊西端、疏勒河中下游，为汉代最西的边关，相传由西域输入和阗玉石取道于此，因而得名。又名小方盘城，与其东北面的当年军粮仓库大方盘城相对。汉初，匈奴控制河西走廊，南结羌人，经常侵扰边疆，并阻绝汉通西域的道路。武帝时战胜匈奴后在河西先后设立酒泉、武威、张掖、敦煌四郡，又筑起屏蔽四郡和走廊通道的长城并置玉门关，以防匈奴。其时约在元封四年（前107）。另一说，初时玉门关并不在此，而在今嘉峪关至赤金峡一带，后来敦煌设郡时移此。今玉门关遗存的城墙尚完整，关城方形如盘，东西长24米，南北宽26.4米，残墙高9.7米，全为黄胶土版筑，面积633平方米，西北各开一门，北门外横过古长城，不及百米即至疏勒河。玉门关隘于阿尔金山之北、疏勒河南岸。汉时气候比现代温和湿润，阿尔金山山麓洪积平原和疏勒河之间当为一片草原或森林草原，汉代"丝绸之路"即经此。当时长城沿线建立许多烽燧亭障，从玉门关沿疏勒河一直向西延伸至今罗布泊北岸。玉门关是通往西域的门户。当时出玉门关至西域的道路有两条，开始均先到古盐泽（今罗布泊）西北的楼兰，然后在此分岔：南道向西南沿昆仑山北麓西行至莎车，出葱岭（今帕米尔高原）至中亚的大月氏、安息；北道经车师前王庭，沿天山南麓西行至疏勒出葱岭至中亚的大宛、康居、奄蔡。西汉初年，匈奴的势力伸展到西域。武帝时汉军攻克匈奴控制的楼兰国，并置西域都护，屯田于乌垒城（今新疆轮台），以保西域通道畅通。汉不但打通了西域通道，而且首次把新疆天山南北地区与内地联为一体。东汉时期，北匈奴以天山南北为基地，不时侵入河西。明帝永平十六年（73）命窦固、耿忠率精骑各出酒泉、敦煌（玉门关）进击北匈奴，再次为开通中亚商路，统一新疆创造了条件。后来班超出使西域也路出玉门关。自魏晋以后，因自安西通新疆东部哈密的道路日见重要，玉门关东移至今安西双塔堡附近。宋以后中国与西方的陆道交通逐渐衰落，玉门关遂废圮。今玉门关故址已为流沙戈壁所侵吞。地下出土许多有价值的文物，为研究汉代边塞屯戍和外交等方面提供了珍贵的资料。

9 司马台长城 1980年，在粉碎"四人帮"，拨乱反正之后，为了加强长城的保护，国务院派出了文化部、国家文物局、公安部、财政部等组成的全国长城破坏情况联合调查组。一个调查组在北京市与河北省交界的古北口附近发现一段长20多公里、形势极为壮观、结构极为奇特的长城。经过历年不断修缮，现已对外开放。这里虽然距北京较远，但由于长城的特色突出，仍然吸引着大量的中外游人。金山岭和司马台这一段长城，是北京市和河北省的交界。由于金山岭首先由河北省滦平县修整开放，司马台则由北京市密云县相继修整开放。且在地形上又有断谷相间隔，因此，归属了北京市和河北省两个省和直辖市管理。金山岭和司马台两处长城，在明朝修筑长城的时候，其军防均属古北口统领，属古北路管辖。这里的长城，虽在明朝初期已有修筑，但均较简陋，现在的城墙、墩台、关隘气势是经抗倭名将戚继光改善加强之后所修筑的。公元16世纪中叶，明穆宗朱载垕为了加强北方的防务，下令将江浙抗倭名将谭伦、戚继光调来北方。委派谭伦为蓟辽总督，戚继光为蓟镇总兵，总理蓟州、昌平一线的防务。明隆庆二年（公元1568年），戚继光将防区内的长城划分为12个路，并调集军民大修长城。由于戚继光奉调北方时，从南方带来了不少士兵。他们为寄托自己的思乡之情，便将江苏省镇江附近金山的名字，冠于这片山上，并把这一带的长城，也呼之为金山岭长城。金山岭司马台是燕山的一条支脉，与燕山主峰雾灵山近在咫尺。在这里千峰

Figure 2: Format of the Corpus

3.2 Training Process and Result Analysis. During the model training process, the same parameter settings were used. The training lasted for one hour and ten minutes, iterating ten times. The final model was obtained at the 11,660th step, where both Training Loss and Validation Loss reached their lowest values. The generated model includes the checkpoint-11660, as illustrated in Figure 3.

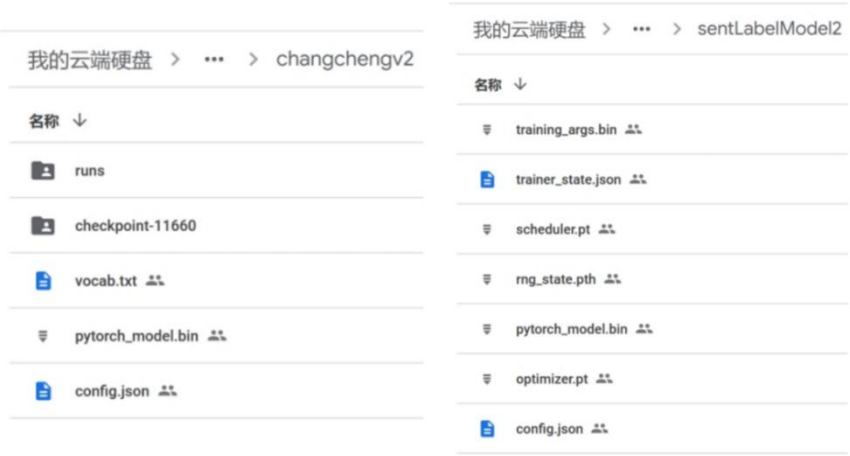


Figure 3: Concept Indexing Model for Great Wall Texts 1

An evaluation was conducted on the accuracy, recall, and F1-score of this model. The evaluation results are presented in the following table.

labels of concepts	Accuracy	Recall	F1
Historical Documents	0.84	0.81	0.78
people	0.78	0.77	0.81
Folk customs	0.65	0.76	0.70

Period	0.79	0.83	0.75
Region	0.75	0.79	0.72
Significance of the Great Wall	0.68	0.66	0.75
Historical Great Wall	0.75	0.79	0.81
Events	0.70	0.69	0.73
Sites	0.75	0.69	0.82
literature	0.75	0.76	0.77
Constructions	0.85	0.80	0.82
Official Positions	0.71	0.79	0.83
average	0.75	0.76	0.77

Table 4: Model Accuracy, Recall, and F-score

The overall concept indexing effect from the final results appears to be unsatisfactory. This could be attributed to several possible factors. Firstly, Some of the training corpus texts are too lengthy, failing to highlight the respective entity's features. While the model, during training, incorporates contextual information into the entity's vector, the learning efficiency for each entity is compromised due to excessively lengthy text. Secondly, the learning effect for the corresponding concept types is relatively poor. When annotating domain concepts that have never been trained, there are issues of annotation errors due to poor learning effects. Concepts like folklore, literature, and events exhibit generally poor performance. Upon analysis, this is primarily due to the similarities in the descriptions between folklore concepts and event concepts. For instance, certain elements within folklore, such as flower performances or cooking methods, overlap with the language forms in some event statements. In literature, most content consists of ancient poems. Poems are known for their concise and condensed language, where each word might represent other types of concepts. For instance, consider the verse 'Qin's unrighteous ways, the seas run dry, Build the Great Wall, hide the northern Huns high' from the poem 'Qi Liang's Wife.' This verse encapsulates concepts related to the 'Qin' dynasty, 'building the Great Wall,' and the 'blocking of the northern barbarians.' Therefore, the annotation performance for this type of concept is inferior compared to others.

The overall training performance in this round is quite poor. Not only is the annotation performance unsatisfactory, but the training also demands a considerable amount of data, thus requiring a longer period. Therefore, it's essential to further analyze and summarize based on the text characteristics of the Great Wall domain books, aiming to further optimize the structure of the corpus.

3.3 Experiment Optimization. Further analyzing the language structure in Great Wall books and summarizing the language structures for each type, we can observe the following linguistic patterns: there are rarely specific definitions provided for concepts like period and region in Great Wall-related books. However, events, figures, and sites often encompass concepts like period and region. Hence, we can consider some concept types in the Great Wall as attribute types. For other non-attribute types,

by analyzing the majority of the book texts, we can deduce and summarize contextual information features within different concept types, as detailed in the table below.

labels of concepts	Language Structure 1	Language Structure 2
Event	(Period, Figure, Region)	(Period, Figure, Region, Event Triggers)
Figure	(Period, Region, Figure)	(Period, Region, Official Position)
Folk Customs	(Region, Period, Figure)	(Region, Period, Folk Custom Triggers)
Site	(Region, Period, Figure, Structure)	
Great Wall through Dynasties	(Period, Structure, Region)	
Literature	(Period, Figure, Region)	

Table 5: Conceptual Language Structure

In this experiment, optimizations were performed on the corpus, implemented specifically in two aspects. Firstly, it involved refining the linguistic structures of the concepts within the corpus. Secondly, efforts were made to minimize the length of the text. Extracting the linguistic structures and semantic relationships from the table above significantly reduced the number of training data. The quantity of corpora for each concept type was reduced by approximately half compared to before. For sentences with similar linguistic structures, only about five were necessary for the model to achieve better learning outcomes. The lesser quantity of corpora for the "Historical Great Wall" category is due to the fewer concepts within this category. Since the "Historical Great Wall" refers to specific long walls built during certain dynasties such as the "Han Great Wall" or "Qin Great Wall," the corpus is relatively limited in comparison. Despite this, compared to training with a large amount of unstructured data in earlier phases, the improvement in performance is noticeable. The table below indicates the current quantity and format of the corpora.

Serial Number	Concept	Quantity
1	Historical Documents	78
2	people	78
3	Folk customs	88
4	Period	85
5	Region	66
6	Significance of the Great Wall	29
7	Historical Great Wall	55
8	Events	66
9	Sites	78

10	Constructions	80
11	Official Positions	85
12	literature	57
Total		845

Table 6: The adjusted number of training corpus

The dataset was divided into training and validation sets following an 8:2 ratio, as depicted in Figure Four. During the model training process, the study employed parameter settings consistent with the previous section. The training lasted for 8 minutes with 10 iterations, obtaining the optimal model at step 2460, where both Training Loss and Validation Loss were minimized. The specific details of the generated model are illustrated in Figure Five.

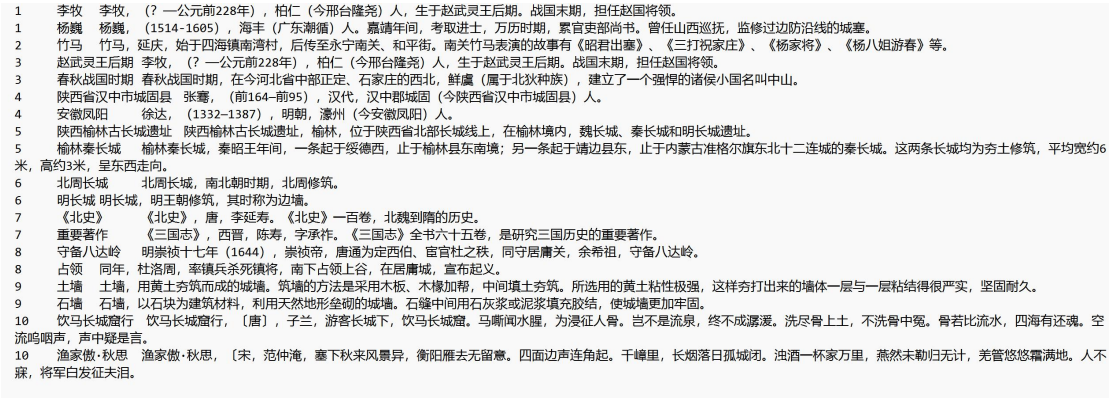


Figure 4 Corpus examples

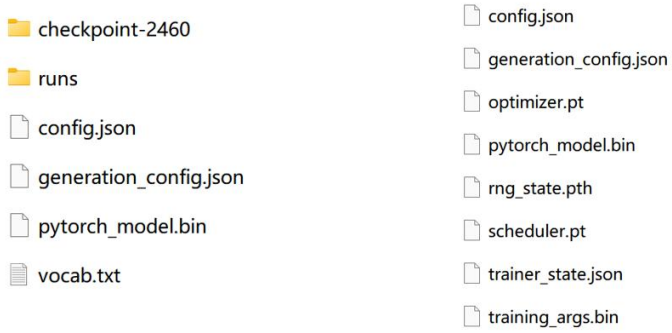


Figure 5 Concept Indexing Model for Great Wall Texts 2

3.4 Experimental results. For the obtained concept indexing results, an evaluation was conducted based on the test set for accuracy, recall, and F1 score, as shown in Table 7.

labels of concepts	Accuracy	Recall	F1
Historical Documents	0.93	0.97	0.95
people	0.85	0.92	0.89
Folk customs	0.98	0.93	0.95

Period	0.95	0.97	0.96
Region	0.97	0.93	0.95
Significance of the Great Wall	0.86	0.80	0.93
Historical Great Wall	0.85	0.79	0.81
Events	0.84	0.91	0.88
Sites	0.92	0.89	0.90
literature	0.95	0.90	0.92
Constructions	0.89	0.83	0.86
Official Positions	0.83	0.88	0.85
average	0.90	0.89	0.90

Table 7: The results obtained by the optimized model

Compared to the previous results, there has been an improvement in the labeling of various concepts. Specifically, "Historical Great Wall" and "Events" have also seen improvements compared to before. However, they still lag behind other concept types, especially "Historical Great Wall," mainly due to its relatively limited corpus, resulting in comparatively poorer model learning effects. The model was uploaded to the parsing platform for parsing entire books. A test was conducted on the structural analysis of 20 selected books, and the specific indexing instance parsing results are presented below:

"Battle of Bai Deng," in the seventh year of Emperor Gao of Han (200 BC), Han's Emperor Liu Bang was besieged by 400,000 elite soldiers under the Xiongnu Chanyu Modu at Bai Deng Mountain, narrowly escaping being captured by the Xiongnu. It marked a notable victory for the Xiongnu over the Han Dynasty.

"Qi Great Wall," remnants of which can still faintly be seen in regions of Shandong, reveals that some places still preserve traces of the wall's remains. It is one of the most preserved sites of the Great Wall from the Spring and Autumn Periods and the Warring States.

"Watchtower" refers to a high platform built on the Great Wall that protrudes from the outer side, used for defending against enemy attacks. In the Song Dynasty's "Wujing Zongyao," it describes the watchtower as a wooden structure projecting outward, built on the face of a horse.

Concept	label	Concept	label	Concept	label
Battle of Bai Deng	Events	Qi Great Wall	Historical Great Wall	Watchtower	Constructions
in the seventh year of Emperor Gao	Period	remnants of which can still be seen in	Region	a high platform built on the Great Wall	Constructions

of Han (200 BC)		regions of Shandong			
Han's Emperor Liu Bang	people	some places still preserve traces of the wall's remains	Constructions	protrudes from the outer side, used for defending against enemy attacks	Constructions
was besieged by 400,000 soldiers at Bai Deng Mountain	Events	preserved sites of the Great Wall from the Spring and Autumn Periods	Significance of the Great Wall	In the Song Dynasty's "Wujing Zongyao	Historical Documents

Table 8: Example of concept annotation results.

Examples of concept annotation results in paragraph text are shown in the table above. Analyzing the concept indexing results, it is found that the granularity of concept indexing is coarser than that of word segmentation and entity, but it covers richer semantics, so the accuracy is also higher. Concept indexing of short sentences can be used for single text. Chapter understanding provides important semantic information, and combined with entity recognition, can better analyze text.

The main causes of errors in concept indexing are concentrated in nested short sentences, such as "There are some locations where the ruins of city walls are still preserved." This short sentence contains both the architectural entity of the city wall and the entity name of the ruins, so the model is Errors occur when predicting short sentences. The results show that the model can identify the syndromes it represents to a certain extent, but due to the limited training corpus, complex semantics of related concepts, and incomplete vocabulary mapping tables, errors in indexing occur. Overall, concept indexing correctly indexes the concepts of most short sentences and can effectively analyze the words and concepts of a single text.

4 Summary and Prospects. Based on the aforementioned, this study focused on concept annotation within the domain of the Great Wall culture. By incorporating prior knowledge, the model's ability for concept annotation was enhanced. This involved integrating prior knowledge into the conceptual indexing model, thereby improving the model's capability to label concepts within the Great Wall cultural texts. Through the utilization of domain pre-trained language models to introduce prior knowledge, the model's adaptation within the Great Wall cultural domain was strengthened. The process involved constructing a conceptual indexing model that fused prior conceptual knowledge. This was achieved by fine-tuning the BERT pre-trained model on domain databases and conducting joint training with the prediction tasks of conceptual entities. This approach aimed to identify known concepts within unstructured text and complete their conceptual categorization. Given that it's impractical to exhaust all training materials within the domain when applying

natural language processing techniques to model training, the study aimed to enhance the efficiency and accuracy of concept indexing within this domain. This was accomplished through an analysis and summary of the language structure and characteristics of Great Wall cultural texts. By leveraging the structural features found in different concepts, a high-quality corpus was constructed to train a more efficient and accurate model.

However, this paper still has some unresolved issues. Firstly, due to the complexity of the Great Wall culture's structure, there often exist intricate structural relationships and semantic connections among various concepts. This leads to occasional errors in the labeling of individual concepts. Additionally, when new concept categories emerge with new semantic structures, the model might encounter issues in concept indexing, resulting in mislabeling or failure to recognize new concepts. Therefore, exploring how to better incorporate semantic structures from prior knowledge into the model remains a challenge and an area for future exploration.

REFERENCES

[1] Li Guanfeng, Zhang Peng. A Web Knowledge Extraction Model Based on Agricultural Ontology [J]. Jiangsu Agricultural Sciences, 2018, 46(4): 201-205.

[2] Wu, Qian. Design and Implementation of an Agricultural Intelligent Question-Answering System Based on Knowledge Graphs [D]. Xiamen: Master's Thesis, Xiamen University, 2019.

[3] ASHISH VASWANI,NOAM SHAZEER,NIKI PARMAR, et al.Attention is all you need [J].arXiv preprint arXiv: 1 706.03 7 62v5,2017.

[4] DAI Z,WANG X,NI P,et al.Named entity recognitionusing BERT BiLSTM CRF for Chinese electronic health records [C],2019:1-5.

[5] ZHANG W,JIANG S,ZHAO S,et al.A BERT-BiLSTM-CRF model for Chinese electronic medical records named entity recognition [C] IEEE,2019:166-169.

[6] Vala H, Jurgens D, Piper A, et a1. Mr. bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts[C] 2015:769-774.

[7] Brooke J, Hammond A, Baldwin T. Bootstrapped text-level named entity recognition for literature[C]20 l 6:344.350.

[8] Xu J, Wen J, Sun X, et a1. A Discourse—Level Named Entity Recognition and Relation Extraction Dataset for Chinese Literature Text[J].Training,l044966(24 65):604.

[9] Xie, Tao. Research and Implementation of Named Entity Recognition Based on Classical Literature[D]. Beijing: Master's Thesis, Beijing University of Posts and Telecommunications, 2018..

[10] Bamman D, Popat S, Shen S. An Annotated Dataset of Literary Entities[C] 2019:2l38-2l44.

[11] Goyal A, Gupta V, Kumar M. Recent named entity recognition and classification techniques: a systematic review[J]. Computer Science Review, 2018, 29: 21-43.