Research Progress and Challenges in Shuishu Informatization: Postprint

Authors: Wu Shuai, Yang Xiuzhang

Date: 2024-07-01T00:00:00+00:00

Abstract

[Purpose/Significance] As a representative of endangered scripts, Shui script has essentially completed its digital preservation under national policy support, yet a gap remains in information construction relative to other pictographic scripts. Identifying these gaps and uncovering deficiencies in Shui script information construction are crucial for advancing the transition of Shui script research toward intelligent application development. [Method/Process] Through field investigation and literature review of Shui script document informatization, this study examines the progression and achievements of Shui script document information construction in China, and explores the associated challenges. [Results/Conclusion] Character-level research on Shui script documents is the most robust, substantially meeting the requirements for ethnic language information construction. However, lexical analysis research remains relatively scarce, and model accuracy requires improvement, attributable to the lack of high-quality Shui script document corpora.

Full Text

Research Progress and Challenges of Informatization Construction of Shui Script Literature

Wu Shuai[1], **Yang Xiuzhang**[2]
(1. College of Information Management, Nanjing Agricultural University, Nanjing 210003, China;
2. School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China)

Abstract:
[**Purpose/Significance**] As a representative of endangered scripts, Shui script has essentially completed its digital preservation with the support of national policies, yet still lags behind other hieroglyphic scripts in informatization

construction. Identifying the gaps between Shui script and other hieroglyphic scripts and recognizing the deficiencies in Shui script informatization are crucial for advancing Shui script research toward intelligent applications. [**Method/Process**] Through field investigations and literature review of Shui script informatization, this paper traces the developmental trajectory and achievements of Shui script literature informatization in China while exploring the associated challenges. [**Result/Conclusion**] The study reveals that research on the "character syntax" of Shui script is the most robust, essentially meeting the requirements for minority language informatization. However, lexical research remains scarce, and model accuracy needs improvement, primarily due to the lack of high-quality Shui script corpora.

Keywords: minority local chronicles; Shui script literature; informatization construction; national literature

Shui script is the collective Chinese translation for the ancient characters of the Shui nationality and their compiled classics. As a "living" hieroglyphic script in the world, the collection, organization, research, and interpretation of Shui script hold significant importance for understanding Shui historical culture, studying primitive religions and social beliefs among ethnic minorities, and exploring the mysteries of ancient Chinese characters. However, the lack of effective organization and information sharing of Shui ancient books and various literature materials has constrained the depth and breadth of Shui cultural research to some extent, indirectly leading to unclear understanding of Chinese Shui studies in the domestic and international academic communities. Moreover, Shui script is primarily recorded on carriers such as embroidery, stone inscriptions, wood carvings, and paper, and is passed down through hand-copying by a small number of Shui masters across generations. Over time, a large number of Shui ancient books and literature materials will continue to be damaged and lost, becoming irreproducible valuable resources. Therefore, as the crystallization of wisdom and valuable wealth derived from Shui cultural development, sorting out the research process of its digital rescue and exploring how to "revitalize" it is a historical mission and a primary measure for inheriting and promoting Shui culture.

In 2000, the National Archives Administration of China officially launched the "Chinese Archival Literature Heritage Project," employing scientific and efficient methods to digitally rescue and protect aging, damaged, and endangered Shui literature. The Central Ethnic Work Conference in 2014 proposed "actively cultivating the consciousness of the Chinese national community," and reiterated in 2019 the importance of "continuously advancing the construction of the Chinese national community," demonstrating the state's emphasis on ethnic minority cultures. How to better protect and research folk ethnic cultures has become a popular research topic. Shui script was included in the first batch of archival literature heritage lists in 2003, and subsequently, the National Archives listed it as a key ethnic ancient book for collection. The State Ethnic Affairs Commission's series on ethnic issues, *Chinese Ethnic Minorities*,

featured the Shui nationality as the first entry. In May 2021, the General Office of the CPC Central Committee and the State Council jointly issued the *14th Five-Year National Archives Development Plan*, explicitly designating the "National Electronic Archives Strategic Backup Center Construction Project" as a key deployment initiative. This increased national attention has raised quality requirements for Shui script digitalization.

1 Current Status of Shui Script Research

Shui script, passed down to the present day, is mainly circulated in the Shui nationality areas of southern Guizhou, with a relatively ancient origin. The earliest recorded study of the Shui nationality in China dates back to 1860, when Mo Youzhi documented Shui script in the preface to *Hongya Ancient Inscriptions Song*. Folklore compilations during the Republic of China period also partially involved Shui script information. Fortunately, although Shui script did not attract scholars' focused attention compared to Zhuang or Uyghur scripts and experienced multiple catastrophes during its inheritance and development, resulting in generational gaps, it continues to survive in the daily lives of Shui people in a living form.

1.1 Limitations of Shui Script Collection and Preservation

Due to property rights awareness and practical needs, numerous scattered folk Shui scripts cannot be collected for professional restoration and proper preservation. Currently, less than one-tenth of the total Shui scripts are housed in libraries or archives. More concerning is that there are currently fewer than one thousand Shui masters nationwide, mostly over 60 years old, and some have passed away without finding successors. Traditional methods of Shui script collection, organization, and research have limitations in manpower, material resources, and technology, making systematic collection, organization, and compilation of large numbers of scattered folk Shui scripts extremely difficult. These methods can no longer meet the new requirements for rescuing endangered ethnic ancient books like Shui script in the information age.

1.2 Four Stages of Shui Script Research

As a rare, complete, and living ancient book of the Shui nationality preserved and used to this day, Shui script has been hailed by experts and scholars as a "living fossil" of world hieroglyphic scripts and was included in the first batch of national intangible cultural heritage lists in 2006. This paper systematically reviews the rescue and protection history of Shui script (see Table 1) and divides it into four stages along a timeline. Before the 1990s (Stage 1), research focused on the origin of Shui culture and basic theoretical sources. During the 1990s (Stage 2), emphasis was placed on the textual exegesis of ancient Shui script and the initiation of Shui script deciphering research. In the 2000s (Stage 3), endangered Shui script rescue was approached from the national memory

level, with initial attempts at Shui script informatization construction. From the 2010s to the present (Stage 4), the pace of Shui ancient book informatization construction has accelerated, enabling the preservation of precious ancient books.

China's organization and research of Shui literature and Shui script began in the 1950s. After the Third Plenary Session of the 11th CPC Central Committee, Shui script rescue and protection work was comprehensively carried out under the care of the Party and governments at all levels, accumulating substantial academic research achievements in areas such as Shui script origin textual research, character exegesis, research overview, and Shui cultural changes. Since the 21st century, domestic and foreign scholars have attached great importance to rescuing endangered Shui ancient books, discovering, collecting, and organizing a large number of Shui literature materials. In addition to traditional research and excavation of Shui script itself, this period also began to consider problems and difficulties in Shui script rescue work and explore various solutions. With the development and popularization of computer technology, Shui script research officially entered the information age, and computer-aided Shui script input, organization, and research methods entered academic 视野, providing substantial academic basis and theoretical foundation for later construction of Shui nationality knowledge graphs, Shui script digital collection research, construction of Shui nationality ontology, and establishment of Shui script electronic databases.

2 Research Progress on Informatization of Shui Nationality Characters

Since the 1980s, computer technology has become the core technology for archival literature protection. However, traditional ethnic scripts, except for Latin alphabets, have not solved the problem of input method implementation, hindering minority research development to some extent.

To examine the research process of Shui nationality character informatization, this paper selects three representative Chinese hieroglyphic scripts: Oracle Bone script, Tibetan script, and Dongba script. Using Oracle archives (Memory of the World International Register [2017]), Tibetan classics (National Precious Ancient Books List [2020]), and Naxi Dongba ancient books (Memory of the World International Register [2003]) as investigation bases, we explore the phased achievements of these three hieroglyphic scripts and compare them with the informatization research process of Shui nationality characters (based on Guizhou Shui script literature), organizing them from perspectives of character research, character digitalization, and digital humanities (see Table 2).

Table 2 compares the informatization research processes of the four hieroglyphic scripts. From the character research perspective, Oracle archives have developed to the stage of exploring article structures, special lexical features, and cross-linguistic comparisons. Tibetan classics have progressed to investigating

character structures, part-of-speech classifications, and literature integration. Naxi Dongba literature has reached the preliminary research stage of character forms, meanings, structures, and pronunciations. Guizhou Shui script literature remains at the entry-level stage of Shui-Chinese translation, Shui script exegesis, character classification, and character-pronunciation induction.

From the digitalization perspective, Oracle archives have achieved intelligent data (Intelligent Data) stage with precise, systematic, and comprehensive Oracle character image databases and computer-automated generation of unrecognized Oracle characters. Tibetan classics have reached linked data (Linked Data) stage using data augmentation to construct special Tibetan character image databases and achieve precise recognition of special handwritten Tibetan through annotation. Naxi Dongba literature has achieved semantic data (Semantic Data) stage through feature extraction and fusion for precise extraction of individual Dongba characters and automatic segmentation of page-level image text lines. Guizhou Shui script literature remains at original data (Original Data) stage, having constructed Shui script input methods and using image enhancement technology to extract Shui ancient book characters in complex environments.

From the digital humanities perspective, Oracle archives have achieved inheritance protection through multidisciplinary exploration of potential historical significance and cultural value, developing Oracle cultural creative products and promotional exhibitions. Tibetan classics have achieved intelligent protection through entity relation extraction using deep learning models and corpus expansion combining Tibetan syntactic structures. Naxi Dongba ancient books have achieved regenerative protection through emotional feature extraction of Dongba paintings using deep learning, construction of Dongba ancient book catalog management systems, and enhanced text representation using word vector conversion. Guizhou Shui script literature remains at original protection stage, using image recognition technology to improve digital extraction and initially attempting to construct Shui script ontology models combining archival literature features.

Comparing the informatization research processes of these four hieroglyphic scripts from the above perspectives, we preliminarily find that Guizhou Shui script literature lacks comprehensive, accurate, and publicly available image-text datasets, with overall research progress lagging behind the other three types. Guizhou Shui script literature's inclusion in the Memory of the World Asia-Pacific Register has attracted scholarly attention to some extent. Therefore, constructing accessible Shui nationality character image datasets and digital archives of Guizhou Shui script literature is the primary task for advancing research. Based on the research development of these four hieroglyphic scripts, we summarize the comparison of their informatization construction status in Table 3 .

3 Research on Character-level Processing of Shui Nationality Characters

Due to the severe lag in the development of Shui nationality ancient characters compared to other scripts, Shui script cannot be studied through manuscript copying like other local chronicles. Therefore, character-level processing of Shui nationality characters is a prerequisite for natural language processing of Shui script and essential for ensuring semantic mining. Since Shui nationality characters have not yet formed standardized character encoding, methods are mainly divided into electronic input based on photoelectric scanning and text input based on character recognition.

3.1 Electronic Input Based on Photoelectric Scanning

Electronic input based on photoelectric scanning refers to the electronic processing of Shui script through scanning and character compilation via text recognition technology. It is divided into explicit segmentation and implicit segmentation based on text recognition methods.

Explicit segmentation constructs a "candidate segmentation-recognition" network path evaluation and optimal path search through single-character classifiers to obtain shape features of individual Shui characters. Currently, the precision rate of commonly used segmentation technologies exceeds 97% [6]. To further improve segmentation accuracy, confidence transformation mechanisms [7] have been introduced on the basis of traditional segmentation technologies. Wang et al. [8] elaborated on the design principles of this method in detail. Kimura et al. [9] used a modified quadratic discriminant function (MQDF) nonlinear classifier for character classification after extracting geometric features from segmentation blocks. However, due to the relatively high complexity of the MQDF model, Liu et al. [10] later combined the nearest prototype classifier (NPC) with MQDF to improve classification speed while maintaining accuracy. Nevertheless, unresolved segmentation issues have constrained the development of explicit segmentation text classification to some extent. With the rapid development of deep learning technology, which has gradually replaced machine learning algorithms, the performance of both single-character and single-sentence recognition has significantly improved.

Implicit segmentation identifies text by presetting candidate characters, using sliding windows to decode between text and sequences to obtain string recognition results. Wang et al. [11] used sliding windows for text encoding, combined with prior probabilities from Hidden Markov Models (HMM) for decoding, and finally selected the optimal posterior probability string through the Viterbi algorithm. The training process of implicit segmentation character recognition does not require character-level annotation, saving substantial annotation time.

3.2 Character Recognition-based Text Input

The emergence of Optical Character Recognition (OCR) technology [12] has made Shui literature informatization more efficient and convenient, improving character feature extraction effects for ancient books [13], engraved texts [14], and natural environments [15]. The implementation of the TH-OCR 2007 [16] ethnic character recognition system promoted Shui character recognition into practical application. However, since Shui characters mostly exist on carriers such as embroidery, stone inscriptions, and wood carvings, many folk Shui scripts have not been digitized. Wang Xiaojuan et al. [17] proposed an image recognition method based on BP neural networks for normalizing handwritten image portions in Shui script. Yang Xiuzhang et al. [18] proposed a Shui character extraction and segmentation algorithm based on adaptive image enhancement and region detection to improve extraction effects. Ding Qiong [19] pioneered the use of Convolutional Neural Networks (CNN) for Shui character recognition, laying the foundation for Shui character recognition systems. Tang Minli et al. [20] attempted to use the Faster-RCNN algorithm for page-level ancient book Shui character recognition, establishing a basis for large-scale digital recognition. Yang Xiuzhang et al. [21] proposed an improved CNN method for ancient character image recognition to address the impact of character shape variations on Shui script recognition. Overall, current research on Shui character processing primarily focuses on character extraction and digitalization of Shui literature, without yet forming database construction for Shui literature.

4 Research on Lexical Analysis of Shui Nationality Characters

Lexical analysis refers to text information processing at the word level and is an important indicator of language intelligence research. Since Shui literature has explicit markers between words, its lexical analysis mainly involves part-of-speech tagging for general words and named entity recognition for special words.

4.1 Research on Part-of-Speech Tagging of Shui Script

Part-of-speech tagging aims to identify the part of speech of unit words. However, natural language processing foundations for Shui nationality characters are weak, and standardized annotated corpora have not yet been formed. Therefore, part-of-speech tagging must be performed with limited corpora. Yang Bei [22] used a small corpus to train an HMM model, combining Viterbi decoding for part-of-speech tagging. Wang Xingjin et al. [23] integrated Shui language rules into the HMM model to further improve tagging performance. Considering sparse word issues, Wang Xingjin et al. [24] used deep learning methods to learn language word-formation features for part-of-speech tagging, though this approach suffers from poor parallelization, long-term information loss, and insufficient feature extraction.

4.2 Research on Named Entity Recognition of Shui Script

Named Entity Recognition (NER) is a special type of part-of-speech tagging in natural language processing, aiming to annotate proper nouns in target corpora to facilitate later domain knowledge graph construction. This is key to transforming Shui script informatization into intelligent applications. This paper categorizes Shui script NER models into character-based, vocabulary-based, and character-vocabulary combined models based on classification granularity.

(1) Vocabulary-based Named Entity Recognition Model
The principle of vocabulary-based Shui script NER models is shown in Figure 1 [Figure 1: see original paper]. Shui script is first translated into Chinese using the *Shui Nationality Characters Chinese Translation Table*, followed by named entity recognition on Chinese words after word segmentation. These methods improve Shui script NER by enhancing Chinese word segmentation effects. Collobert and Weston [25] used word embeddings to replace traditional handcrafted features, employing CNN to extract Shui script features combined with Conditional Random Field (CRF) models for entity category prediction. Ma et al. [26] fused Bidirectional Long Short-Term Memory (Bi-LSTM) and CNN to build a Bi-LSTM-CNN model for improved semantic feature extraction. Chen et al. [27] improved Bi-LSTM-CRF model performance in identifying entity word boundaries by enhancing segmentation effects.

(2) Character-based Named Entity Recognition Model
The principle of character-based Shui script NER models is shown in Figure 2 [Figure 2: see original paper]. Shui script is first translated into Chinese using the *Shui Nationality Characters Chinese Translation Table*, followed by named entity recognition on Chinese characters, effectively avoiding error propagation caused by difficult entity boundary identification. Dong et al. [28], as the first domestic team to use radical sets for character-based NER, pioneered research pathways for Shui script NER. Zhu and Wang [29] considered dependencies between characters, attempting to use Bidirectional Gated Recurrent Unit (Bi-GRU) for sentence-level global information acquisition and incorporating attention mechanisms for local feature weighting to initially achieve semantic fusion. Gu et al. [30] innovatively attempted to capture character rule features using two regular modules, constructing orthogonal spaces to fuse captured features for local semantic enhancement.

(3) Character-Vocabulary Combined Named Entity Recognition Model
Character-vocabulary combined NER models enhance vector semantic expression capabilities by incorporating lexical semantic understanding on the basis of character-based models. Ghaddar et al. [31] pioneered a lattice LSTM model that uses dictionary matching to achieve character semantic enhancement on the LSTM foundation, with the specific model principle shown in Figure 3 [Figure 3: see original paper]. Sui et al. [32] constructed a collaborative graph network model of word-character inclusion graphs, transition graphs,

and character lattice graphs to effectively solve information loss in lattice structures. Gui et al. [33] integrated lexical features and attention mechanisms into graph neural networks to alleviate lexical conflicts and resolve linguistic ambiguity.

With further development of artificial intelligence technology, the Transformer framework has entered the historical stage, driving advances in image recognition and natural language processing. Xue et al. [34] added relative position encoding attention mechanisms to the Transformer framework to enhance dependency relationships between characters and words. Li et al. [35] integrated the Transformer framework into the Lattice LSTM framework to achieve information interaction between characters and potential words, efficiently solving lexical conflict problems.

As Transformer matures, pre-trained models dominated by this architecture far outperform static embedding models based on vocabulary-character combinations in NER tasks. Liu et al. [36] used special encoding markers to identify word boundaries in sentences before inputting them into pre-trained models, effectively improving recognition performance.

5 Challenges and Recommendations for Shui Script Informatization Construction

Through literature review, we find that current Shui script informatization construction significantly lags behind other hieroglyphic scripts, with its core still remaining at the digital storage level and facing numerous challenges in informatization construction and intelligent applications.

5.1 Challenges Faced by Shui Script Informatization Construction

Based on achieved results, we identify the following main challenges in Shui script informatization and intelligent application research:

(1) Severe Corpus Scarcity
Shui script corpus construction involves corpus selection, collection, processing, and analysis. China's Shui script corpus construction began in the 1990s, but was severely limited by the scarce basic theoretical research and inefficient archival digitalization at that time, affecting both scale and quality. Moreover, Shui script lacks a 健全 and standardized annotation scheme, relying solely on Shui masters for annotation, which severely constrains the progression of Shui script informatization toward intelligent applications.

(2) Predominance of Informal Texts
Shui script has relatively few classic materials, mostly consisting of daily life records, with scarce official formal documents. The uneven distribution of corpus data types is unfavorable for accurate model recognition. Moreover, the largest proportion consists of real-time records by Shui masters and some hand-

written copies, which feature complex diversity and chaotic text semantics, increasing the difficulty of text informatization processing.

(3) Severe Carrier Damage
Due to the special nature of Shui nationality characters, Shui script is only passed down through hand-copying by a small number of Shui masters over 60 years old. Apart from these hand-copied versions, Shui script carriers are mostly original materials such as classics, embroidery, stone inscriptions, wood carvings, and gold carvings. Over time, severe oxidation and corrosion seriously hinder digital extraction of Shui script.

5.2 Recommendations for Shui Script Informatization Construction

To address these challenges, we propose the following five recommendations for Shui script informatization construction:

(1) Automated Annotation of Shui Script
Automated annotation is the foundation for transforming Shui script informatization toward intelligent applications and can effectively alleviate researchers' data annotation pressure. However, pre-trained model training data does not include Shui literature data. Increasing Shui annotation data in pre-trained models to achieve automated annotation is an urgent problem for Shui database construction.

(2) Fine-grained Mining of Shui Script
Improving text corpus granularity is key to accurately obtaining knowledge through text mining techniques. Since Shui script contains ten fine-grained entity types including astronomy, marriage, sacrificial rituals, and geography, with significant professional knowledge differences between entity types, deep fine-grained division is necessary to provide a good data foundation for intelligent applications.

(3) Construction of Shui Script Feature Representation
As a hieroglyphic script, Shui nationality characters have radicals in their forms and vowels/consonants in their pronunciations, which are beneficial for entity word boundary division. Therefore, feature representation methods conforming to Shui script's word-formation rules can be constructed by combining these characteristics to achieve natural language processing tasks in its unique domain.

(4) Construction of Standardized Annotation Systems
Transfer learning models can achieve natural language processing research across different languages of the same type or different types of the same language, already realized in some Chinese natural language processing datasets. This provides a theoretical foundation for Shui script technology research. However, cross-language transfer learning suffers from label matching and domain difference issues. Therefore, applying transfer learning models to the Shui script domain requires early construction of standardized annotation systems.

(5) Lightweight Pre-trained Models

The Transformer architecture has been well-applied in large-scale corpora, and pre-trained models have achieved good cross-domain and cross-language transfer learning. However, due to significant differences between Shui script and pre-trained model language features, fine-tuning requires substantial computing power for iterative text semantic learning, causing resource waste. Balancing model recognition accuracy and lightweight design is essential for Shui script intelligent applications.

Through literature review, we find that with national policy support, Shui nationality characters have achieved network transmission based on unified encoding, laying the foundation for text informatization and digital protection of endangered Shui script. However, a significant gap remains compared to other hieroglyphic script research. Currently, "character syntax" research on Shui script is the most robust, basically meeting minority language informatization needs. However, "lexical" research is relatively scarce, primarily due to severe corpus scarcity, predominance of informal texts, and severe carrier damage. Considering the small market share of Shui script informatization, which cannot attract industry attention, and the lack of standardized informatization construction standards, the uneven ontology construction levels in academic collection, organization, and analysis of endangered Shui script hinder the accuracy of downstream text mining tasks and impede the development of Shui script intelligent applications.

We believe that applying low-resource information processing technologies such as unsupervised, multi-task, few-shot, and zero-shot learning to Shui script informatization construction and subsequent intelligent applications is key to overcoming the challenges in Shui script informatization.

References:

[1] Liu Xiaocheng, Wu Liyan. Communication of Ethnic Policies from the Perspective of the Chinese National Community: Connotation, Function and Framework [J]. *News History*, 2022(5): 75-82.
[2] Wang Yuying, Zhang Zhijie, Li Nianfeng. Research on the Construction of Information Processing Models and Countermeasures for Traditional Culture of Ethnic Minorities in the Big Data Era [J]. *Information Science*, 2022, 40(7): 154-160, 168.
[3] Qu Zhilin. Analysis of the Compilation Status of Shui Script Archives [J]. *Lantai World*, 2016(1): 25-27.
[4] Li Minghua. Proposal on Establishing a National Electronic Archives Strategic Backup Center [J]. *China Archives*, 2022(3): 20.
[5] Meng Yaoyuan. Review and Reflection on Ten Years of Shui Script Rescue and Protection Work [J]. *Literature and History Expo (Theory)*, 2016(1): 23-26.
[6] Yang W, Jin L, Tao D, et al. Drop Sample: A new training method to enhance deep convolutional neural networks for large-scale unconstrained handwritten Chinese character recognition [J]. *Pattern Recognition*, 2016, 58: 190-203.
[7] Liu C L. Classifier combination based on confidence transformation [J]. *Pattern Recognition*, 2005, 38(1): 11-28.

[8] Wang D H, Liu C L. Learning confidence transformation for handwritten Chinese text recognition [J]. *International Journal on Document Analysis and Recognition*, 2014, 17(3): 205-219.
[9] Kimura F, Takashina K, Tsuruoka S, et al. Modified quadratic discriminant functions and the application to Chinese character recognition [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1987 (1): 149-153.
[10] Liu C L, Nakagawa M. Evaluation of prototype learning algorithms for nearest-neighbor classifier in application to handwritten character recognition [J]. *Pattern Recognition*, 2001, 34(3): 601-615.
[11] Wang Z R, Du J, Wang J M. Writer-aware CNN for parsimonious HMM-based offline handwritten Chinese text recognition [J]. *Pattern Recognition*, 2020, 100: 107102.
[12] Mori S, Nishida H, Yamada H. *Optical character recognition* [M]. John Wiley & Sons, Inc., 1999.
[13] Xue Chunhan, Jin Xiaofeng. A Recognition Method for Few-Shot Korean Ancient Books Characters Based on Transfer Learning [J]. *Journal of Yanbian University (Natural Science Edition)*, 2021, 47(4): 350-355.
[14] Renqing Dongzhu. Research on Tibetan Ancient Books Woodblock Character Recognition Based on Deep Learning [D]. Lhasa: Tibet University, 2021.
[15] Hong Song, Gao Dingguo, Sanpai Cairang, et al. Detection and Recognition of Uchen Tibetan Script in Natural Scenes [J]. *Computer Systems & Applications*, 2021, 30(12): 332-338.
[16] Qian Lihua. Successful Development of Multi-Ethnic Character Document Recognition System on Unified Platform [N]. *China Ethnic News*, 2007-01-30(001).
[17] Wang Xiaojuan, Bai Yanping. Research on Handwritten Digit Recognition Method Based on BP Neural Network [J]. *Mathematics in Practice and Theory*, 2014, 44(7): 112-116.
[18] Yang Xiuzhang, Wu Shuai, Xia Huan, et al. Research on Shui Nationality Character Extraction and Recognition Based on Adaptive Image Enhancement Technology [J]. *Computer Science*, 2021, 48(S1): 74-79.
[19] Ding Qiong. Research on Implementation of Shui Script Character Feature Extraction and Classification Method on Matlab Platform [J]. *Electronic Technology & Software Engineering*, 2020(14): 155-157.
[20] Tang Minli, Xie Shaomin, Liu Xiangrong. Detection and Recognition of Handwritten Characters in Shui Script Ancient Books Based on Faster-RCNN [J]. *Journal of Xiamen University (Natural Science)*, 2022, 61(2): 272-277.
[21] Yang Xiuzhang, Shi Yi, Li Na, et al. An Improved Convolutional Neural Network Method for Arabic Character Image Recognition [J]. *Information Technology and Informatization*, 2021(9): 6-11.
[22] Yang Bei. Research on Lao Word Segmentation and Part-of-Speech Tagging Methods [D]. Kunming: Kunming University of Science and Technology, 2016.
[23] Wang Xingjin, Zhou Lanjiang, Zhang Jinpeng, et al. Research on Semi-Supervised Lao Part-of-Speech Tagging Integrating Word Prediction [J]. *Journal of Chinese Computer Systems*, 2019, 40(12): 2500-2505.

[24] Wang Xingjin, Zhou Lanjiang, Zhang Jian'an, et al. Multi-Task Lao Part-of-Speech Tagging Method Integrating Word Structure Features [J]. *Journal of Chinese Information Processing*, 2019, 33(11): 39-45.
[25] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning [C]//*Proceedings of the 25th International Conference on Machine Learning*. 2008: 160-167.
[26] Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF [C]//*Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 2016: 1064-1074.
[27] Chen X, Qiu X, Zhu C, et al. Long short-term memory neural networks for Chinese word segmentation [C]// *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015: 1197-1206.
[28] Dong C, Zhang J, Zong C, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition [M]// *Natural Language Understanding and Intelligent Applications*. Springer, Cham, 2016: 239-250.
[29] Zhu Y, Wang G. CAN-NER: Convolutional attention network for Chinese named entity recognition [C]// *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2019: 3384-3393.
[30] Gu Y, Qu X, Wang Z, et al. Delving deep into regularity: A simple but effective method for Chinese named entity recognition [C]// *Proceedings of the NAACL-HLT (Findings)*. Seattle: Association for Computational Linguistics, 2022.
[31] Ghaddar A, Langlais P, Rashid A, et al. Context-aware adversarial training for name regularity bias in named entity recognition [J]. *Transactions of the Association for Computational Linguistics*, 2021, 9: 586-604.
[32] Sui D, Chen Y, Liu K, et al. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network [C]// *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019: 3830-3840.
[33] Gui T, Ma R, Zhang Q, et al. CNN-based Chinese NER with lexicon rethinking [C]// *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. Macao: IJCAI, 2019: 4982-4988.
[34] Xue M, Yu B, Liu T, et al. Porous lattice transformer encoder for Chinese NER [C]// *Proceedings of the 28th International Conference on Computation Linguistics*. Barcelona: International Committee on Computational Linguistics. 2020: 3831-3841.
[35] Li X, Yan H, Qiu X, et al. FLAT: Chinese NER using flat-lattice transformer [C]// *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics. 2020.
[36] Liu W, Fu X, Zhang Y, et al. Lexicon enhanced Chinese sequence labeling using BERT adapter [C]// *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 2021: 5847-5858.