On the Development of Interlinearized Ancient Literature of Ethnic Minorities: A Case Study of the Interlinearization of Ancient Written Tibetan Literature

Congjun Long and **Bo An**
Institute of Ethnology and Anthropology,Chinese Academy of Social Sciences
Building 6, Zhongguancun Nandajie 27, Beijing, China
{lcj,anbo}@cass.org.cn

Abstract

Ancient ethnic documents are essential to China's ancient literature and an indispensable civilizational achievement of Chinese culture. However, few research teams are involved due to language and script literacy limitations. To address these issues, this paper proposes an interlinearized annotation strategy for ancient ethnic literature. This strategy aims to alleviate text literacy difficulties, encourage interdisciplinary researchers to participate in studying ancient ethnic literature and improve the efficiency of ancient ethnic literature development. The interlinearized annotation consists of original, word segmentation, Latin, annotated, and translation lines. In this paper, we take ancient Tibetan literature as an example to explore the interlinearized annotation strategy. However, manually building a large-scale corpus is challenging. To build a large-scale interlinearized dataset, we propose a multi-task learning-based interlinearized annotation method, which can generate interlinearized annotation lines based on the original line. Experimental results show that after training on about 10,000 sentences (lines) of data, our model achieves 70.9% and 63.2% F1 values on the segmentation lines and annotated lines, respectively, and 18.7% BLEU on the translation lines. It dramatically enhances the efficiency of data annotation, effectively speeds up interlinearized annotation, and reduces the workload of manual annotation.

1 Introduction

The excellent traditional culture of ethnicity is an essential part of Chinese culture, an important cultural heritage of the Chinese nation, and a valuable asset to human civilization. Many excellent traditional cultures have been recorded in ancient ethnic literature (Bender, 2015), some of which record the process of creating the great history of the Chinese nation together and the vivid facts of exchanges and interactions among various ethnic groups. They contain rich national unity and progress ideas and are necessary historical resources for witnessing the Community of the Chinese Nation (Meng et al., 2023). Therefore, the in-depth excavation of ancient ethnic literature is conducive to promoting traditional Chinese culture and showing the historical events of the formation of Sense of Community for the Chinese Nation (Long et al., 2023).

China is rich in ancient ethnic literature, but studying ancient ethnic literature faces many difficulties. First, the degree of digitization is relatively low due to the lack of public resources; second, limited by language and script literacy, the group of ancient ethnic literature research and utilization is small. In exploring the formation of Chinese civilization and promoting Chinese culture, how more disciplines and researchers pay attention to, study, develop, utilize, and popularize the excellent traditional culture contained in ancient ethnic books is an issue worth exploring. The General Office of the CPC Central Committee and the General Office of the State Council issued the Opinions on Promoting the Work of Ancient Books in the New Era, emphasizing the need to encourage interdisciplinary research methods. The 'text structuring', 'knowledge systemization' and 'intelligent utilization' of ancient books are actively carried out (Lei et al., 2022).

The documentary properties and unique cultural attributes of minority antiquarian literature have made it a focus of interdisciplinary experts and a laboratory for interdisciplinary research (Long et al., 2023). However, constructing most ancient ethnic literature re-

Proceedings of the Ancient Language Processing Workshop associated with RANLP-2023, pages 222–231,
held in Varna Bulgaria, Sept 8, 2023.
https://doi.org/10.26615/978-954-452-087-8_025

sources is still difficult to meet the needs of multiple disciplines. For example, experts in computational linguistics focus on the information processing of ancient ethnic documents and need a cooked corpus with information on word segmentation, annotation, entity recognition, and translation, and then carry out deep text mining. Experts in the field of library intelligence explore the collection, collation, cataloging, and citation of multi-language ethnic ancient texts from the perspective of knowledge organization and knowledge management of ancient texts, and build catalog search libraries and full-text search libraries to serve readers better. Linguistics researchers are concerned about the phonology, vocabulary, and grammar of the national languages in the multi-language ancient ethnic document to assist in the construction of the ancient Chinese phonetic system, analysis and comparison of the Chinese and the people's language relations, to explore the language homology differentiation clues, summarize the phonetic, lexical and grammatical type characteristics and the evolutionary path of the language. Researchers in history focus on historical elements such as time, place, people, and events in the ancient texts of multi-language ethnic groups and explore the political systems, economic systems, social histories, and foreign exchanges of different ethnic groups. Scholars in ethnic culture explore the traditional culture, folk customs, cultural heritage, and traditional handicrafts recorded in the ancient texts of multilingual ethnic groups. Experts in religion, philosophy, art, and traditional medicine also hope to obtain the knowledge they need from multilingual ethnic literature.

To better meet the needs of multidisciplinary utilization of ancient ethnic texts, this paper proposes a strategy of interlinearized annotation of ethnic ancient texts, converting the content of ethnic ancient texts into five lines of data, namely, the original line (original line), the line of folk language sub-word (segmentation line), the line of Latin alphabet transcription (transcription line), the line of grammar annotation (annotation line) and the line of Chinese meaning translation (translation line). Researchers from different disciplines can use these annotations to analyze and study the literature content. In conducting interlinearized annotation research, manual annotation was mainly used in the early stage. With the accumulation of annotation data, this paper proposes a multi-tasking framework based on deep learning to automatically generate interlinearized annotation data to assist manual annotation and finally build a large-scale textual structured database of ancient ethnic literature to lay the data foundation for further development and utilization of ancient ethnic documents in multiple disciplines.

2 Related Work

China is a multi-ethnic country, and in the long history of the formation and development of the Chinese nation, people of all ethnic groups have shared honor and disgrace and are closely related to each other, creating Chinese civilization and culture together. The multilingual chapter-aligned, sentence-aligned, and word-aligned historical documents handed down or excavated archaeologically are the best proof of the exchange and intermingling of people from all ethnic groups.

Among the Chinese and Tibetan bilingual aligned historical documents, chapter-aligned documents are the most numerous, followed by sentence-aligned and less word-aligned documents. Chapter-aligned documents are both translated from Chinese into ethnic texts, such as the four ancient Tibetan translations of the Shang Shu Zhou Shu (Wong, 2016) in the first collection of the 1978 Paris photocopy of the Selected Tibetan Documents of the Bibliothèque Nationale de French; there are also translations from ethnic texts into Chinese, such as the oath on the west side of the Tang-Fan Alliance monument, which is a Chinese-Tibetan aligned sentence pair (Li F G, 2007). Most of the materials in the form of word control are found in the dictionary category and word list categories, such as the Great Collection of Translation Nominalities (Z, 2013), the Dunhuang Tibetan texts P.T. 1257 and P.T. 1261 (X, 2014), and the Imperial Five-Style Qing Wenjian (Q, 2000). However, there are not many materials on the full-text word alignment of ancient literature texts. Scholars of linguistics who study ancient ethnic literature often have to translate the documents.

In general, the source and translated texts are still aligned in terms of chapter alignment, such as the translation of Baxie (Supplementary Text) (Ba S N, 1990), the Chinese translation of the History of Buddhism in Buton (Bu D, 2007), and the Tibetan King's Tale (Suo N J Z, 2002), among others. However, some scholars have also adopted the word-alignment model, such as the Study of Tibetan Fatwas in the 8th-9th centuries (Z, 2007), etc. Linguistic researchers have been more rigorous in organizing documentary materials, especially in dialectal and ethnolinguistic materials, and have mostly adopted the word alignment model. This paradigm is used for language text materials in the Chinese Ethnolinguistic Compendium Series, the Newly Discovered Languages of China Series, and the Endangered Languages of China Series. For example, an example of the annotated text for the language of the security language on page 1918 of Languages of China.

The German publishing house Lincom GmbH has been funding the publication of interlinearly annotated corpora and scholarly works in minor languages worldwide for many years. Tikaram Poudel published Rajbanshi Grammar and Interlinearized Text (an Indo-Aryan language of Nepal and Bengal) in 2006 (Poudel, 2006); Karnakhar Khatiwada published A Reference Grammar in 2017 of Dhimal (King, 2008) describing writings and text annotation Interlinearized texts in Dhimal with Grammar notes (Khatiwada, 2017) (interlinearized annotated texts in Dhimal). To date, the publisher has published more than 500 works in small languages. Sino-Tibetan linguists Randy J. LaPolla & Dory Poa also published Rawang Texts grammatically annotated texts at Lincom Europa (LaPolla and Poa, 2001).

Computer experts have developed interlinearized annotation tools to assist linguists in advancing interlinearized annotation successfully. The American Standard Interchange Language (SIL) organization has developed Toolbox [1] annotation tool; British scholars have developed Eudico Linguistic Annotator (ELAN) annotation tool [2], and French scholars adopted the Interlinear Text Editor software (ITE) technology. Chinese scholars have used Toolbox tool to organize and publish the series 'Grammatical Annotated Texts of Chinese Ethnic Languages' (D, 2016), which is a total of 20 books covering 20 languages or dialects of five principal language families or groups in China, namely Tibetan-Burmese, Miao-Yao, Dong-Tai, South Asian and Altaic, with a total word count of about 10 million words. These software tools are widely used in the linguistic community, making it easier and faster for linguists to annotate the corpus and enhancing the standardization of corpus annotation. However, the common drawback is that they mainly rely on manual operations and fail to introduce natural language processing techniques for low-resource languages, especially natural language information techniques.

Interlinearized annotation is similar to the goal of word alignment in machine translation, where the word alignment technique is to obtain word boundaries in sentence pairs and achieve translation alignment based on bilingual pairs, which is a core task in machine translation (Bahdanau et al., 2014). However, the research results devoted to word alignment methods belong to the early stage of statistical machine translation. With the development of ethnolinguistic information processing and the promotion of the 'One Belt, One Road' strategy, machine translation of low-resource languages has become a popular research topic (Ranathunga et al., 2023), and some research results discussing the word alignment between Chinese and Mandarin have appeared. For example, Zhao Yang and Zhou Long discussed the Min-Chinese scarce resource neural machine translation technique (Zhao Yang, 2019); (Su L Y, 2018) discussed the word alignment method in Mongolian-Chinese machine translation; (Liu J M, 2011). Studied the Han-Vi word alignment. However, the current machine translation commonly adopts deep neural network technology, which does not need to discuss word alignment methods separately.

In recent years, the concepts of 'exploring the origin of Chinese civilization' and 'forging a sense of Chinese national community' have been proposed, and interdisciplinary fields have jointly focused on transcribed texts of

[1] https://software.sil.org/software-products/
[2] http://sites.bu.edu/elsa/elan-coding/

ethnic minority oral discourse and ancient ethnic literature. The increasing demand for interlinearized annotation of ethnic texts has made interlinearized annotation a research paradigm. However, it is challenging to meet the needs of multidisciplinary and multilingual interlinearized annotated corpus by manual annotation. This paper combines the linguistic fine-grained annotation paradigm and multi-tasking techniques to conduct automatic interlinearized annotation.

3 Ancient Ethnic Literature Interlinearized Annotation

3.1 The Format of Ancient Literature Interlinearized Annotation

The target of interlinearized annotation is the full text of ancient ethnic literature without explicit markers between words, which need to be divided into 'words' for the full text, and then convert the traditional ethnic script into Latin transcription or international phonetic symbols by word. The words are translated into the other language, and the function words are marked with their grammatical function. The grammatical functions are labeled with English abbreviation tags common to the linguistic community. The final output consists of a line of the original language, a Latin transcription or an International Phonetic Alphabet line corresponding to the line of the minor language, and a line of the translation marker. Under the current technical conditions, meaningful translation lines cannot be obtained automatically and need to be translated manually. Figure 1 takes Tibetan as an example to show an example of interlinearized annotation.

The original line is the original text of the ethnic literature. The segment line is a unit of words (partly morphemes and phrases), a 'word' or 'word suffix' for the input text. However, 'word' is the most basic unit for the machine to understand the text. To satisfy the deep analysis and mining of the text, marking the word boundary is the most basic task. The materials of ancient texts marked with word boundaries help users understand ancient texts. They can be used for training to develop automatic lexical analysis tools for ancient texts, providing resources to support the information processing of ancient literature.

The transcription lines are ethnic texts transcribed in Latin alphabet or the international phonetic alphabet. The aim is to create a cross-reference database of ethnic scripts in syllables. The annotation line: the meaningful words in the annotated lines are translated into Chinese or English. Function words are marked with the abbreviated form of their grammatical function in English, and the abbreviated form often uses a combination of capital letters, which comes from the English word 'AGENT' and semantically indicates the administration of things. This internationally accepted grammatical mark is conducive to the dissemination of national ancient literature materials to the world. The translation line is the Chinese or English translation of the original line.

3.2 The Schema of Ancient Literature Interlinearized Annotation

Designing symbols for interlinearized annotation requires consideration of the usage needs of ancient ethnic literature, which have been discussed earlier. For the same ancient ethnic literature, different researchers have different needs. Such as syllogisms, word segmentation, and named entity recognition are common needs for researchers. Semantic annotation is a common need for linguistics-related disciplines such as syntax and language information services. Named entity annotation (NER) is a common need for history and literature, etc. Chinese translation of meaningful words is closely related to machine translation. This paper focuses on syntactic and semantic annotation and entity annotation, whose annotation materials can meet the needs of most disciplines. Grammatical and semantic annotation can reflect functional words' grammatical meaning and semantic function. Entity annotation includes proper names such as person, place, and time. Labels for function words are generally composed of two or three letters, taking the first three letters of the English word. If repetition is encountered, the abbreviated letters are modified as appropriate. When a grammatical function requires more than one English word to be represented, the appropriate combination of letters from multiple words is selected. The combination of labels also fol-

Type	Content
Origin Line	དེ་བློན་པོས་རྗེའི་སྙན་དུ་གསོལ།
Segment Line	དེ་ བློན་པོ/ས་ རྗེ/འི་ སྙན་ དུ་ གསོལ་ པ།
Latin Line	de blon po s rje vi snyan du gsol pas
Annotation Line	DEM 大臣 AGE 王 GEN 耳朵 ALL 禀报 CLC
Translated Line	大臣把那（情况）禀报到国王的耳朵里……

Figure 1: The example of Tibetan interlinearized annotation.

lows specific rules, and the linguistic community usually adopts the Leipzig Terminology Rules (The Leipzig Glossing Rules) system [3], which was jointly developed by the Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology and the Department of Linguistics of the University of Leipzig. In our study of interlinearized annotation of ancient civil texts, the Leipzig labeling system was employed as the primary basis, with the addition of some labels. The tagging system includes person and number, grammatical, tense, tone, mood, demonstrative, special word classes, syntax, noun-pronoun correlative markers. The grammatical tags employed for interlinear annotation will vary with the degree of refinement of the corpus, and the tags listed here are only the main ones. Some minority language scripts require an extension of the tagging system according to specific needs but keep the basic system unchanged. The NLP-based NER is adopted.

4 Human-computer Interaction Interlinearized Annotation Platform

Corpus annotation is time-consuming and labor-intensive; however, annotated corpora can provide essential resources for ancient literature research and are indispensable. With enough training corpus, NLP algorithms can assist corpus annotation, such as NER, relation extraction, text classification, and machine translation. The corpus annotation is also a common task in NLP. The interlinearized annotation of ancient ethnic literature is a pioneering work, and by manually annotating a certain scale of training data, the NLP algorithms can assist in data annotation.

To advance the research in this paper, we have developed a semi-automated interlinearized annotation platform. The interlinearized annotation platform mainly has the following modules.

(1) Interlinearized annotation operation module. This module completes the task of interlinear annotation in the original language of ancient ethnic literature and accomplishes four main functions:

(a) Automatic conversion from the original language to the transcription line. The conversion from the original Chinese text to the Latin alphabet depends on the mapping table between the original syllables and the Latin syllables. For example, the Tibetan to-Latin conversion collects about 28,000 syllables, including modern Tibetan, ancient Tibetan, Sanskrit syllables transcribed in Tibetan characters and punctuation marks. Also, it includes syllables that conform to Tibetan spelling rules but do not exist in existing literature.

(b) Automatic word segmentation. The interlinearized annotation is based on the annotation of words (or phrases), and word segmentation is a necessary process.

(c) Annotation line including meaningful word translations and grammatical labels. To be compatible with various text editors and convenient for different researchers, the interlinear annotation results are stored as XML format files, with a set of brackets {} to indicate the four levels of interlinear corresponding lexical entries, which are filled in with the morphological analysis before the split form, Latin transcription, morphological analysis after the Latin transcription and annotation information, respectively.

(d) Manual proofreading. Manual proofreading needs to do three checking aspects: filling in vacant paraphrases, correcting paraphrase errors due to multiple meanings, and annotation errors due to subtext errors.

[3] https://www.eva.mpg.de/lingua/resources/glossing-rules.php

(e) Batch export and import of annotation data: to reduce the workload of manual annotation, the import and export of data are supported after a certain number of interlinear annotation data are completed.

(f) Interlinearized annotation corpus retrieval module. This module supports structured retrieval and can meet the needs of general literature information retrieval, such as the retrieval of Latin transcriptions of ethnic scripts, pairs of translated words and labels; the extraction of cross-referenced word lists, the extraction of named entities, and other functions.

5 Automatic Interlinear Annotation Method based on Multi-task Learning

5.1 Method

In this paper, we propose NLP method to promote the interlinearized annotation of ancient ethnic literature for deep analysis and exploration. In the previous work, we have annotated a dataset of ancient Tibetan interlinearized annotation, so we take ancient Tibetan as an example to introduce the automatic generation method of interlinearized annotation based on deep learning, and the interlinearized annotation of other ethnic ancient literature is similar to ancient Tibetan. The interlinearized annotation dataset of ancient Tibetan includes original, segmentation, Latin, annotated, and translation lines. Among them, the segmentation lines are obtained by automatic word segmentation based on the original lines (Liu H D, 2012), the Latin lines can be transcribed directly by the Tibetan-Latin conversion table, and the annotation lines are generated based on the segmentation lines with corresponding meaningful words and function words annotated. The original line, the syllogism line, and the annotated line are all valuable for generating translation lines. Therefore, the ancient Tibetan interlinearized annotation model is mainly used to generate segmentation, annotated, and translation lines. Translation lines are translations from Tibetan sentences to Chinese sentences, which belong to the research scope of machine translation. Regarding the current research base, the generation of Chinese lines is more complex and requires human intervention. The model's input is the original line, and the output is the content of the segmentation line, the annotation line and the translation line. The information of the annotated line depends on the content of the segmentation line, and the information of the translation line depends on the information of both the segmentation line and the annotated line. Therefore, the pipeline model (Li et al., 2020) is employed in modeling, and its architecture is shown in Figure 4-a. The pipeline approach consists of three models: (1) the word segmentation model: the input of this model is the original line, and the output is the result of the word segmentation line; (2) the annotation model: the input of this model is the information of the original line and the word segmentation line, and the output is the result of the annotation line; (3) the translation model: the input of this model is the original line, the word segmentation line and the annotation line, and the output is the result of the translation line. The pipeline method splices the outputs and inputs of different models, such as using the output of the segmentation line as the input of the annotation line, which facilitates the implementation of the model. However, the pipeline method suffers from problems such as error propagation. For example, the segmentation model can only utilize the information of the original text line, but the information of the annotation line also has important value for the segmentation model, and the error of the segmentation line may cause the obvious error of the annotation line information. However, since the models are independent, the error information cannot be effectively transferred to the segmentation model, and this part of the information cannot be utilized.

Recently, multi-task learning models replaced pipeline-based models in several fields, such as segmentation and annotation models, named entity recognition and entity linking models (Nguyen and Grishman, 2015). The advantage of multi-task learning models is that they can make full use of the correlation between different tasks, e.g., there is a strong correlation between the Tibetan word segmentation and lexical annotation tasks, and the results of word segmentation determine the

text block boundaries of linguistic annotation. In contrast, the results of lexical annotation can, in turn, verify whether there are errors in the word segmentation results. Therefore, multi-task learning approaches are widely applied due to the advantages in modeling multiple related tasks. Interlinearized annotation requires the generation of corresponding segmentation lines, annotated lines, and translation lines based on the original text lines, a typical multiple-related task suitable for modeling using a multi-task learning framework.

Based on the above analysis, this paper designs a multi-task model to conduct the word segmentation, annotation, and translation models. The model's input is the original text line, and the shared coding layer encodes the input information. Then, different upper-layer models are used to model the output tasks (word segmentation, annotation, and translation lines). The word segmentation model, the annotation model, and the translation model share the embedding layer (Embedding) (Lai et al., 2016) and the encoding layer (Bi-directional Long Short-Term Memory, BiLSTM) (An et al., 2018; An and Long, 2021), and the word segmentation line is based on the original line, and the words are segmented from each other using spaces. Therefore, in this paper, the word segmentation is modeled as a sequential annotation task, and the task layer uses Conditional Random Field (Sutton et al., 2012). The annotation line contains grammatical annotation information and word translation information for a sequence generation task, which is modeled as an encoder-decoder sequence generation task in this paper. The translation line is the translation of the original line content into Chinese, which is a typical machine translation task, and is also modeled as an encoder-decoder sequence generation task in this paper (Guo et al., 2019).

5.2 Experimental Settings

To verify the effectiveness of the interlinearized annotation model, we completed four interlinearized annotated ancient Tibetan literature, Baxie (Ba S N, 1990), Weixie, Zhumian Shi (Bu D, 2007), and Di Wu Shiji (Schneider, 2002), through the annotation platform. The dataset consists of 12,284 sentences, and we divide it into the training set, development set, and test set according to the scale of 8:1:1.

5.3 Experimental Settings

This section describes the implementation framework and hyperparameters employed in the experiments. We utilize Pytorch to implement a multi-task model. The dimension of Tibetan syllables is 100; the coding layer is BiLSTM. We employ CRF to implement word segmentation tasks. In the annotation task, we decode the sequences using a single-layer BiLSTM to generate grammatical function labels. We use a single-layer BiLSTM to generate the translation line in the translation task. This paper employs the pipeline-based model as the baseline, including the BiLSTM+CRF-based word segmentation model, BiLSTM+BiLSTM-based annotation model, and encoder-decoder-based translation model. These models are trained separately, with input and output data transfer and information interaction. This paper employs Precision, Recall, and F1-value to evaluate the results of the segmentation line and annotation line. We employ BLEU (Papineni et al., 2002) to evaluate the result of the translation line.

5.4 Experimental Result

We conduct three sets of experiments: the first set of experiments adopts a multi-task learning approach, aiming to implement interlinearized annotation of ancient Tibetan literature, generating segmentation lines, annotated lines, and translation lines based on the original lines; the second set of experiments is a stripping experiment, using a multi-task learning approach to model segmentation lines and annotated lines, segmentation lines and translation lines, and annotated lines and translation lines, respectively, to analyze the effect of joint learning of different tasks; the third set of experiments utilizes a pipeline model as a baseline model to compare the performance of proposed models.

5.5 Experimental Results

The input of this experiment is the original text line, and the output includes the segmentation line (Seg), the annotation line (Ann),

and the translation line (Tra). The experimental results are shown in Table 1. The experimental results show that the multi-task learning-based model (Multi) proposed in this paper significantly improves all of the three tasks of segmentation lines, annotation lines, and translation lines (6.7%, 15.6% and 32.6%, respectively). The multi-task learning-based model achieved better performance than the pipeline model (Pipe). Therefore, the multi-task-based model can achieve better performance in interlinearized annotation task.

Task	Model	P	R	F	BLEU
Seg	Mult	74.2	67.8	70.9	-
	Pipe	68.2	64.7	66.4	-
Ann	Mult	66.2	60.4	63.2	-
	Pipe	54.1	55.2	54.6	-
Tra	Mult	-	-	-	18.7
	Pipe	-	-	-	14.1

Table 1: The result of multi-task learning model.

5.6 Ablation Study

5.7 Ablation Study

The ablation study aims to analyze the effect of different task combinations. In this experiment, group A (segmentation model + annotation model), group B (segmentation model + translation model), and group C (annotation model + translation model). Table 2 shows the results of the three experimental groups of experiments. Based on the experimental results, we can draw the following conclusions: (a) In both groups of multi-task learning in which the segmentation model participates, the segmentation result improves (F1 value 66.4%), indicating that the results of both the annotated lines and the translation lines can improve the performance of the segmentation model; (b) The segmentation model achieves better results in the results of group A, indicating that the annotation line provides better feedback to the segmentation model than the translation line; (c) In group A, the multi-task model achieves better results than the pipeline model, indicating that the annotation model can give effective feedback to the segmentation model. However, the performance of the annotated rows in group C experiments has a significant decrease, indicating that the information of the segmentation rows has an essential impact on the results of the annotated line; (d) The results of translation lines in both groups B and C decreased compared with the multi-task learning model but better than the pipeline-based model, indicating that the information of both segmentation lines and annotated lines has auxiliary value for the translation model.

Task	Model	P	R	F	BLEU
A	Seg	72.1	66.2	69.0	-
	Ann	60.3	59.4	59.8	-
B	Seg	71.8	66.0	68.8	-
	Tra	-	-	-	15.3
C	Ann	50.3	48.2	49.2	-
	Tra	-	-	-	14.7

Table 2: The result of ablation study.

6 Conclusion

In this paper, we discuss the idea and vision of interlinearized annotation of ancient ethnic literature from the perspective of data resource normalization and sharing. Taking ancient Tibetan literature as an example, we propose accumulating corpus based on manual interlinearized annotation and then using machine learning to conduct automatic annotation. This research provides a new research paradigm for developing and utilizing ancient ethnic literature in China, especially the structured data of interlinearized annotation, which lays a good foundation for ancient literature development and utilization. In the future, we plan to construct ancient language knowledge based on an interlinearized annotation dataset.

7 Acknowledgments

This work is supported by the Natural Science Foundation of China (22BTQ010), the National Natural Science Foundation of China (62076233) and the Innovation Project major research of Chinese Academy of Social Sciences (2022MZSQN001).

References

Bo An, Bo Chen, Xianpei Han, and Le Sun. 2018. Accurate text-enhanced knowledge graph representation learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 745–755.

Bo An and Congjun Long. 2021. Neural dependency parser for tibetan sentences. *Transactions on Asian and Low-Resource Language Information Processing*, 20(2):1–16.

Tong J H Ba S N. 1990. *Ba Xie*, volume 10. Sichuan Ethnic Publishing House.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Mark Bender. 2015. Ethnic minority literature. *A Companion to Modern Chinese Literature*, pages 261–275.

Pi W C Bu D. 2007. *History of Budun Buddhism*, volume 9. Gansu Ethnic Publishing House.

Jiang D. 2016. Chinese national language grammar annotation text series. *(No Title)*.

Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *Ieee Access*, 7:63373–63394.

Karnakhar Khatiwada. 2017. *Interlinearized Texts in Dhimal with Grammar Notes*. Lincom Europa.

John Timothy King. 2008. *A grammar of Dhimal*. Ph.D. thesis, Leiden University.

Siwei Lai, Kang Liu, Shizhu He, and Jun Zhao. 2016. How to generate a good word embedding. *IEEE Intelligent Systems*, 31(6):5–14.

Randy J LaPolla and Dory Poa. 2001. *Rawang texts*, volume 18. Lincom Europa.

Yuying Lei, Xilong Hou, Xiaoguang Wang, et al. 2022. The logic and approach of digital reconstruction of ancient books in the data intelligence. *Digital Human Research*, 2(2):46.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Ke W N. W Q L Li F G. 2007. *Study on Ancient Tibet*, volume 1. Beijing Tsinghua University Press.

Zhao W N Liu H D, Nuo M H. 2012. Segt: A practical tibetan word segmentation system. *Journal of Chinese Information Processing*, 26(1):97–103.

Ai S Liu J M, Tu R G. 2011. Research on statistical machine translation-based chinese-uyghur word alignment. *Computer Applications and Software*, 28(4):57–59.

Congjun Long, Bo An, and Shengyan Zhang. 2023. Research on the construction of knowledge graph of old tibetan inscriptions. *Library And Information Service*, 67(8):141.

Linglei Meng, Jingnan Xing, and Mingyan Tan. 2023. Exploration of cultivating a sense of community for the chinese nation in" data structures" course teaching. *International Journal of Education and Humanities*, 7(3):136–139.

Thien Huu Nguyen and Ralph Grishman. 2015. Relation extraction: Perspective from convolutional neural networks. In *Proceedings of the 1st workshop on vector space modeling for natural language processing*, pages 39–48.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Tikaram Poudel. 2006. *Rajbanshi grammar and interlinearized text*, volume 34. Lincom.

Jiang Q. 2000. A textual research on the compilation of qing wenjian in imperial four and five styles.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Hanna Schneider. 2002. Tibetan legal documents of south-western tibet: structure and style. In *Proceedings of the Ninth Seminar of the IATS, 2000. Volume 1: Tibet, Past and Present*, pages 415–427. Brill.

Niu X H Su L Y, Zhao Y P. 2018. The study on ethnic-to-chinese scare-resource neural machine translation. *Journal of Chinese Information Processing*, 32(6):44–51.

Liu Q L Suo N J Z. 2002. *Tibet Wangtongji*, volume 2. Nationalities Publishing House.

Charles Sutton, Andrew McCallum, et al. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Sun-tik Wong. 2016. The critical study of political pbligation in zhou shu of shang shu and liji=" shang shu, zhou shu" ji" li ji" 'zheng zhi yi wu'zhi yan jiu. *HKU Theses Online (HKUTO)*.

Dang Z Z X. 2014. Ancient tibetan dictionary.

Huang W Z. 2007. *Research on Tibetan vows in the 8th and 9th centuries*, volume 8. Nationalities Publishing House.

Zhang Y Z. 2013. A collection of translated names – the origin of tibetan bilingual dictionary.

Wang Qian etc Zhao Yang, Zhou Long. 2019. The study on ethnic-to-chinese scare-resource neural machine translation. *Journal of Jiangxi Normal University*, 43(6):630–637.