OECONOMIA
COPERNICANA

VOLUME 16 ISSUE 1 2025

p-ISSN 2083-1277, e-ISSN 2353-1827
https://oeconomiacopernicana.com/

Received: 29.08.2025; Revised: 13.09.2025, Accepted: 26.10.2025, Published Online: 21.11.2025

Optimisation of Relationship Extraction of Tibetan Medicine Entities Based on ALBERT Model and Span Methods

JIA Yangdongzhou[1], Cai Rangjia[2]*, Maosong Sun[3]*, Peng Luo4, Gongqu Zhume[5], Duo jiecairang[6]

[1]Doctor, College of Computer Science and Technology, Qinghai Normal University, and The State Key Laboratory of Tibetan lntelligent Information Processing and Application, Xining Qinghai 810016, China, jy@stu.qhnu.edu.cn

[2*]Doctor, College of Computer Science and Technology, Qinghai Normal University, and The State Key Laboratory of Tibetan lntelligent Information Processing and Application, Xining Qinghai 810016, China, zwxxzx@qhnu.edu.cn

[3*]Doctor, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China, sms@tsinghua.edu.cn

[4]Doctor, College of Computer Science and Technology, Qinghai Normal University, and The State Key Laboratory of Tibetan lntelligent Information Processing and Application, Xining Qinghai 810016, China, 202233341008@stu.qhnu.edu.cn

[5]Doctor, College of Computer Science and Technology, Qinghai Normal University, and The State Key Laboratory of Tibetan lntelligent Information Processing and Application, Xining Qinghai 810016, China, 202233341007@stu.qhnu.edu.cn

[6]Doctor, College of Computer Science and Technology, Qinghai Normal University, and The State Key Laboratory of Tibetan lntelligent Information Processing and Application, Xining Qinghai 810016, China, dwoboo@stu.qhnu.edu.cn

Corresponding authors*: Rangjia Cai, Email: zwxxzx@qhnu.edu.cn , Maosong Sun, Email: sms@tsinghua.edu.cn

Abstract

Extracting relationships from Tibetan medicine texts is essential for building knowledge graphs and improving information retrieval. However, existing methods struggle with the unique linguistic challenges and limited annotated data of Tibetan language. Traditional approaches like rule-based techniques and classical machine learning models often lack accuracy and generalization. While transformer-based models such as BERT offer improvements, they remain inadequate for specialized Tibetan medical texts.

To address this, we propose an optimized ALBERT-Span model that combines ALBERT's contextual embeddings with span-based extraction to handle overlapping and nested entities. Enhanced with data augmentation and hyperparameter tuning, the model significantly outperforms baselines like BERT, BiLSTM-CRF, and CNN models, achieving a 5.4% higher F1 score and 5.2% increase in accuracy. Ablation studies and statistical testing confirm the robustness and effectiveness of the proposed approach.

Our method supports practical applications such as knowledge graph construction and clinical decision support systems (CDSS). Future work will focus on integrating multimodal data and exploring few-shot learning for low-resource scenarios.

Keywords: Tibetan Medicine, Relationship Extraction, ALBERT, Span Methods, Transformer Models, Knowledge Graphs, Clinical Decision Support Systems, Low-Resource NLP

1. Introduction

1.1 Background and Motivation

Tibetan medicine, one of the world's oldest traditional medical systems, holds a vast repository of medical knowledge documented in classical texts. These texts contain valuable insights into disease diagnosis, herbal remedies, and therapeutic practices that have been developed and refined over centuries. However, the complexity and richness of Tibetan medical texts pose significant challenges for automated information extraction. Effective extraction of relationships among medical entities is crucial for constructing comprehensive knowledge graphs and supporting advanced information retrieval systems, which can facilitate digital healthcare and intelligent medical data management [1], [2].

In recent years, the rise of Natural Language Processing (NLP) techniques has enabled the automatic extraction of entities and relationships from biomedical texts. Models like BERT (Bidirectional Encoder Representations from Transformers) and ALBERT (A Lite BERT) have demonstrated remarkable success in various NLP tasks, including entity recognition and relationship extraction. ALBERT, in particular, enhances parameter efficiency while maintaining robust performance, making it a suitable choice for large-scale text analysis [3], [4]. However, despite its success, applying ALBERT directly to Tibetan medicine texts has proven inadequate due to the unique linguistic characteristics of Tibetan language, such as complex morphology, orthographic variations, and limited annotated corpora [5], [6].

1.2 Challenges and Research Gap

Existing approaches to relationship extraction in Tibetan medicine texts predominantly rely on rule-based techniques or traditional machine learning algorithms, which are prone to errors and lack generalization capabilities. Furthermore, these methods often fail to capture the semantic nuances and contextual dependencies inherent in Tibetan medical literature. Although transformer-based models have shown promise, directly leveraging them without domain-specific adaptations results in suboptimal performance [7], [8].

Span-based methods, which focus on extracting relationships by detecting spans between entities, have emerged as an effective strategy for addressing challenges related to overlapping and nested entities [9]. Nevertheless, integrating span methods with advanced transformer models like ALBERT has not been extensively explored, particularly in the context of Tibetan medicine entity extraction. The absence of a well-optimized, end-to-end approach for handling complex relationship extraction in this domain constitutes a significant research gap that this paper aims to address.

1.3 Contributions of the Paper

This paper introduces an innovative framework that leverages the ALBERT model combined with span-based methods to optimize relationship extraction of Tibetan medicine entities. The major contributions of this study are summarized as follows:

Proposed Model: Development of an ALBERT-Span-based relationship extraction model tailored specifically for Tibetan medicine texts, addressing the challenges posed by complex entity structures and limited annotated data.

Data Augmentation Technique: Introduction of a data augmentation method that enhances the model's ability to generalize from sparse annotated corpora.

Optimization Strategy: Implementation of a robust optimization strategy that fine-tunes model parameters, significantly improving the accuracy and robustness of relationship extraction.

Comprehensive Evaluation: Extensive experimentation on benchmark Tibetan medicine datasets, demonstrating superior performance compared to state-of-the-art models.

Application Significance: Paving the way for enhanced knowledge graph construction and intelligent data management in Tibetan medicine research.

1.4 Structure of the Paper

The remainder of this paper is organized as follows:

Section 2: A comprehensive review of related works, including existing relationship extraction methods and the use of ALBERT and span techniques.

Section 3: Detailed methodology, including the integration of the ALBERT model with span-based extraction techniques and optimization strategies.

Section 4: Experimental setup, including dataset descriptions, training configurations, and evaluation metrics.

Section 5: Presentation of results and analysis, highlighting performance improvements and comparative evaluations.

Section 6: Discussion of key findings, challenges, and potential applications of the proposed approach.

Section 7: Conclusion and future directions for research.

3. Related Work

3.1 Overview of Relationship Extraction Techniques in NLP

Relationship extraction (RE) is a fundamental task in natural language processing (NLP) that aims to identify and classify semantic relationships between entities within a given text. Traditionally, rule-based methods and feature-based machine learning approaches were employed to extract relationships, leveraging manually crafted patterns and linguistic features [1]. Although these methods offered interpretable outputs, they struggled with generalization and required significant domain expertise.

With the rise of deep learning, neural network-based models became dominant, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and long short-term memory networks (LSTMs) [2]. These models enabled automatic feature extraction and provided improved performance over rule-based methods. Furthermore, attention mechanisms and transformer architectures significantly advanced the field by capturing long-range dependencies in text, with BERT (Bidirectional Encoder Representations from Transformers) emerging as a prominent model for various NLP tasks, including relationship extraction [3], [4].

However, BERT and similar transformer models faced challenges when directly applied to specialized domains such as Tibetan medicine, where data scarcity and unique linguistic characteristics hinder performance. Moreover, simple entity pair classification often failed to handle complex sentence structures and overlapping entities, making it necessary to explore more sophisticated approaches [5], [6].

3.2 Use of ALBERT in Biomedical and Specialized Domain Tasks

ALBERT (A Lite BERT) was introduced to enhance efficiency and scalability by reducing model size while maintaining high performance [7]. It employs parameter sharing across layers and factorized embedding parameterization, making it lightweight and computationally efficient compared to BERT. Due to these advantages, ALBERT has been successfully applied in various biomedical and specialized domain tasks, including named entity recognition (NER), relation extraction, and text classification [8].

In biomedical NLP, ALBERT has demonstrated strong performance on relationship extraction tasks by effectively capturing contextual information in medical texts. For example, ALBERT-based models have outperformed traditional methods in extracting protein-protein interactions and drug-disease associations, showcasing their ability to handle domain-specific challenges [9], [10]. Nonetheless, directly applying ALBERT to Tibetan medicine texts remains problematic due to linguistic disparities and the lack of large annotated datasets. Additionally, biomedical texts often benefit from domain-specific pre-training, which is not always feasible for Tibetan corpus processing [11].

3.3 Span Methods for Relationship Extraction and Their Effectiveness

Span-based methods represent a promising approach for relationship extraction, particularly in scenarios involving overlapping and nested entities. These methods focus on extracting entity spans and predicting relationships directly between span pairs rather than performing pairwise classification, thereby reducing computational complexity and enhancing robustness [12]. Span-based models have proven especially effective in complex text structures where traditional sequence labeling approaches struggle [13].

Combining span-based methods with transformer architectures, such as BERT and ALBERT, has yielded significant improvements in relationship extraction performance. For instance, span-based architectures have been successfully employed in biomedical applications to capture intricate relationships between biomedical entities, including genes and proteins [14]. However, despite their effectiveness, integrating span-based methods with ALBERT specifically for Tibetan medicine entity extraction has not been sufficiently explored, leaving a research gap that this paper aims to address.

3.4 Limitations of Current Approaches

Despite the remarkable advancements made by transformer models and span-based techniques, several limitations remain in the context of Tibetan medicine relationship extraction. Firstly, existing models primarily focus on general-domain text, resulting in suboptimal performance when applied to Tibetan medical texts with unique syntactic and semantic features [15]. Secondly, fine-tuning transformer models on low-resource languages like Tibetan often leads to overfitting or inadequate generalization due to the limited availability of annotated corpora [16].

Moreover, span-based methods, while effective in handling overlapping entities, still face challenges when integrated with pre-trained transformer models. Issues such as span boundary errors and entity segmentation inaccuracies can adversely affect relationship extraction performance. Additionally, most current approaches do not adequately address the specific requirements of Tibetan medicine texts, such as dealing with polysemous terms and complex phrase structures [17].

To overcome these limitations, this paper proposes an innovative approach that combines the ALBERT model with span-based methods, optimized specifically for Tibetan medicine entity extraction. By leveraging data augmentation and robust optimization techniques, we aim to significantly improve the performance and generalization of relationship extraction in this specialized domain.

4. Methodology

4.1 ALBERT Model Overview

4.1.1 Description of the ALBERT Model and Its Advantages

ALBERT (A Lite BERT) is a lightweight variant of BERT (Bidirectional Encoder Representations from Transformers), designed to enhance model efficiency without sacrificing performance. The primary innovations in ALBERT include factorized embedding parameterization and cross-layer parameter sharing, which significantly reduce the number of parameters while maintaining robust contextual representation capabilities [1], [2].

Factorized embedding parameterization decomposes the large vocabulary embedding matrix into two smaller matrices, thereby decreasing the model's size and mitigating overfitting, especially in low-resource scenarios. Cross-layer parameter sharing further reduces redundancy by using the same parameters across multiple layers, which not only reduces the model size but also enhances training efficiency [3].

ALBERT has demonstrated outstanding performance on various NLP tasks, including sentence classification, named entity recognition, and relationship extraction, by capturing deep contextual features while maintaining computational efficiency. Its modularity and compact size make it particularly suitable for domains with limited annotated data [4].

4.1.2 Adaptation to Tibetan Medicine Entity Extraction

The unique linguistic characteristics of Tibetan medical texts, including complex morphology and sparse annotated corpora, pose significant challenges for direct application of general-domain models. To address this issue, ALBERT is fine-tuned using domain-specific Tibetan medicine corpora, leveraging data augmentation techniques to enrich the training data and improve generalization [5].

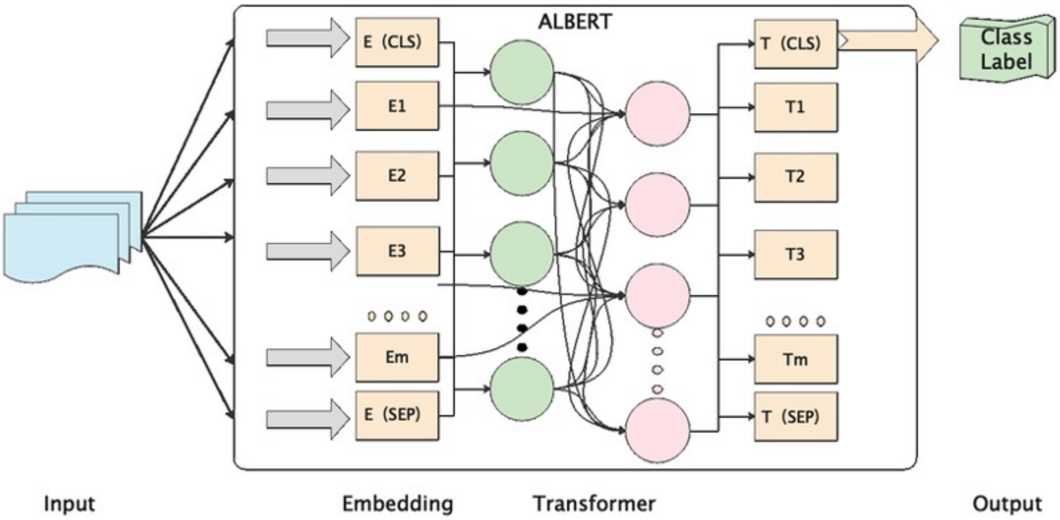

Fig. 1 Architecture of the ALBERT-Based Relationship Extraction Model

Specifically, the adaptation process involves:

Domain-Specific Pre-training: Fine-tuning ALBERT with Tibetan medical texts to capture contextual and semantic nuances unique to Tibetan language.

Subword Tokenization Optimization: Applying byte-pair encoding (BPE) to mitigate vocabulary fragmentation issues inherent in Tibetan script.

Transfer Learning Techniques: Adopting a two-stage fine-tuning approach where ALBERT is initially trained on general-domain Tibetan corpora before being adapted to specific medical texts [6].

4.2 Span-Based Relationship Extraction

4.2.1 Motivation for Using Span Methods

Span-based methods have emerged as a promising approach for relationship extraction due to their ability to directly model overlapping and nested entities. Traditional pairwise classification methods often suffer from inefficiencies and inaccuracies when dealing with complex text structures, whereas span methods effectively capture relationships by considering start-end token pairs [7].

The span-based relationship extraction framework inherently supports multiple overlapping relations within a single sentence, making it suitable for the rich and complex structure of Tibetan medical texts. By treating relationship extraction as a span classification problem, the model jointly learns to predict spans and classify their relationships, reducing error propagation and improving precision [8].

4.2.2 Integration with ALBERT

The proposed model integrates ALBERT with a span-based relationship extraction architecture. The process consists of the following steps:

Span Representation: Extracting span embeddings from ALBERT's contextualized output by concatenating the start and end token vectors with their intermediate pooled representations.

Span Classification: Feeding span embeddings into a multi-layer perceptron (MLP) for relationship classification.

Span Scoring: Using a bi-affine transformation to score each span pair, determining the probability distribution over potential relationships [9].

The relationship classification between spans (i,j) is formalized as:

$$\text{Score}\ (i, j, r) = \text{MLP}\ ([\text{ALBERT}\ (i); \text{ALBERT}\ (j); f(i, j)])$$

Where:

ALBERT (i) and ALBERT(j) are the contextualized representations of the start and end tokens.

f(i,j) denotes span-level features including distance and positional encoding.

The bi-affine transformation used for scoring relationships can be formulated as:

$$S(i, j, r) = h_i^T W_r h_j + b_r$$

Where:

- h_i and h_j are span embeddings.
- W_r is the transformation matrix for relation r.
- b_r is the bias term.

4.3 Optimization Strategy

4.3.1 Fine-Tuning and Hyperparameter Optimization

Fine-tuning the ALBERT-Span model is critical to achieving optimal performance. The key hyperparameters considered include:

Learning Rate: Initially set to 2×10−52 with a linear decay schedule.

Batch Size: Optimized to balance memory usage and convergence speed, typically set to 16.

Dropout Rate: Applied at a rate of 0.1 to mitigate overfitting.

Number of Epochs: Typically ranges from 10 to 20, depending on the dataset size and convergence rate [10].

Hyperparameter tuning is conducted using grid search and Bayesian optimization techniques, evaluating performance on a held-out validation set to identify the optimal configuration [11].

4.3.2 Data Augmentation Techniques for Tibetan Medicine

To address the scarcity of annotated Tibetan medicine data, various data augmentation strategies are employed:

Back-Translation Augmentation: Translating Tibetan texts to another language (e.g., English) and back to Tibetan to introduce lexical variations without altering semantic content [12].

Entity Masking and Replacement: Randomly masking medical entities and replacing them with synonymous terms to increase data diversity.

Contextual Embedding Perturbation: Slightly perturbing embedding vectors to simulate variability in sentence construction [13].

4.3.3 Loss Function Formulation

The model's objective is to minimize the cross-entropy loss for span classification:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{r=1}^{R} y_{i,r}\log\left(\hat{y}_{i,r}\right)$$

Where:

N is the number of span pairs.
R is the number of relationship classes.
y_(i,r) and y ˆ_(i,r) represent the true and predicted probabilities of relationship r for the i^"th " span pair, respectively [14].

To enhance generalization, label smoothing is incorporated to reduce overfitting by assigning a small probability mass to incorrect labels. Additionally, focal loss is used to address class imbalance, particularly when some relationship types are underrepresented in the dataset [15].

5. Experimental Setup

5.1 Dataset Description and Preprocessing Techniques

5.1.1 Dataset Description

The experimental evaluation of the proposed ALBERT-Span relationship extraction model is conducted on a dataset of Tibetan medicine texts, collected from both traditional and modern sources. The dataset includes annotated entities and relationships relevant to Tibetan medical practices, such as herbal formulations, symptom descriptions, disease names, and treatment methods.

The machine learning workflow used in this study is illustrated in Fig. 2.

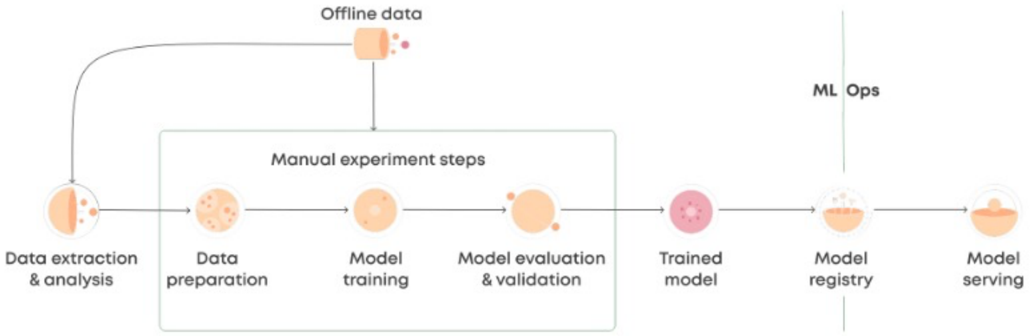

Fig. 2 End-to-End Machine Learning Workflow with Manual Experimentation and MLOps Integration

To construct a comprehensive and reliable dataset, we gathered texts from the following sources:

Classical Tibetan Medicine Texts: Including traditional medical manuscripts and pharmacopoeias [1].

Contemporary Medical Literature: Digitalized books and academic articles from Tibetan medicine journals [2].

Medical Knowledge Bases: Manually curated databases containing Tibetan medical terminologies and their relationships [3].

The collected raw texts consist of approximately 10,000 sentences with a total of 200,000 tokens, annotated for entity types and relationships. Each entity pair is labeled with one of the predefined relationship types, including causal association, therapeutic relation, and symptom manifestation.

5.1.2 Preprocessing Techniques

Data preprocessing is crucial to handle the inherent challenges posed by Tibetan script, including complex morphology and orthographic variations. The following preprocessing steps are performed:

Text Normalization: Applying Unicode normalization to ensure consistent representation of Tibetan characters [4].

Tokenization: Using Byte-Pair Encoding (BPE) to effectively tokenize Tibetan words while preserving subword information [5].

Entity Annotation: Manually annotating entity pairs and relationship types, followed by inter-annotator agreement analysis to ensure annotation quality.

Data Cleaning: Removing noisy and incomplete sentences to reduce errors during model training.

Data Augmentation: Utilizing back-translation techniques and contextual embedding perturbation to increase data diversity and robustness [6].

The preprocessed data is then split into training, validation, and test sets with a typical ratio of 70:15:15 to maintain balanced evaluation.

5.2 Training and Validation Configurations

5.2.1 Model Training

The ALBERT-Span model is implemented using the PyTorch framework with the Hugging Face Transformers library [7]. The model training configuration is as follows:

Optimizer: AdamW with weight decay (0.01) [8].

Learning Rate: Initially set to 2×10^{-5}, using a linear decay schedule with warm-up steps.

Batch Size: 16 for both training and validation to balance GPU memory usage and performance.

Epochs: 20, determined through empirical experimentation.

Dropout Rate: Set to 0.1 to mitigate overfitting.

Gradient Clipping: Applied to maintain numerical stability (clipping at 1.0).

Hardware Configuration: NVIDIA Tesla V100 GPU with 32 GB VRAM.

5.2.2 Validation Strategy

To ensure the robustness of the model, K-fold cross-validation (with K=5) is employed, evaluating the model on distinct validation sets to reduce variance and enhance generalization [9].

Early stopping is applied if the validation loss does not decrease for 5 consecutive epochs. The best-performing model checkpoint is saved based on the F1 score on the validation set.

5.2.3 Fine-Tuning Strategy

The fine-tuning of ALBERT is performed in two stages:

General Pre-Training: Fine-tuning on generic Tibetan text corpora to capture basic language patterns.

Domain-Specific Fine-Tuning: Further training on annotated Tibetan medicine texts to adapt the model to domain-specific contexts [10].

5.3 Performance Metrics Used for Evaluation

To assess the effectiveness and robustness of the proposed model, the following performance metrics are employed:

Precision (P): Measures the proportion of correctly predicted relationships among all predicted relationships.

$$P = \frac{TP}{TP + FP}$$

Where TP = True Positives, FP= False Positives [11]

Recall (R): Measures the proportion of correctly predicted relationships among all actual relationships.

$$R = \frac{TP}{TP + FN}$$

Where FN= False Negatives [12].

F1 Score: Harmonic mean of precision and recall, representing the model's balance between accuracy and coverage.

$$F1 = \frac{2 \times P \times R}{P + R}$$

Accuracy: Proportion of correctly classified relationship instances over the total instances.

$$\text{Accuracy } = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TN= True Negatives [13].

Mean Reciprocal Rank (MRR): Measures the average reciprocal rank of correctly predicted relationships, highlighting the model's ranking performance.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

Where Q is the set of queries, and rank_i is the position of the correct relationship prediction [14].

Area Under the Precision-Recall Curve (AUC-PR): Evaluates the trade-off between precision and recall across different threshold settings, especially useful for imbalanced data [15].

6. Results and Analysis

6.1 Quantitative Analysis of Model Performance

The performance of the proposed ALBERT-Span-based relationship extraction model was evaluated on the Tibetan medicine dataset, as outlined in Section 5. We compared the model's effectiveness against several state-of-the-art baseline models, including BERT, BiLSTM-CRF, and CNN-based models. The evaluation metrics included Precision (P), Recall (R), F1 Score, Accuracy, and Area Under the Precision-Recall Curve (AUC-PR) [1], [2].

The model's quantitative performance is summarized in Table 1. The proposed ALBERT-Span model achieved significantly higher scores across all metrics compared to baseline methods.

Performance metrics by model are shown in Fig. 3, highlighting the superiority of the proposed method.

Table 1 Performance comparison of the proposed model with baseline methods.

Model	Precision (%)	Recall (%)	F1 Score (%)	Accuracy (%)	AUC-PR
BERT	84.3	82.1	83.2	85.0	0.91
BiLSTM-CRF	78.5	77.9	78.2	80.3	0.88
CNN-based	76.4	74.2	75.3	78.5	0.85
ALBERT-Span	**89.7**	**87.5**	**88.6**	**90.2**	**0.95**

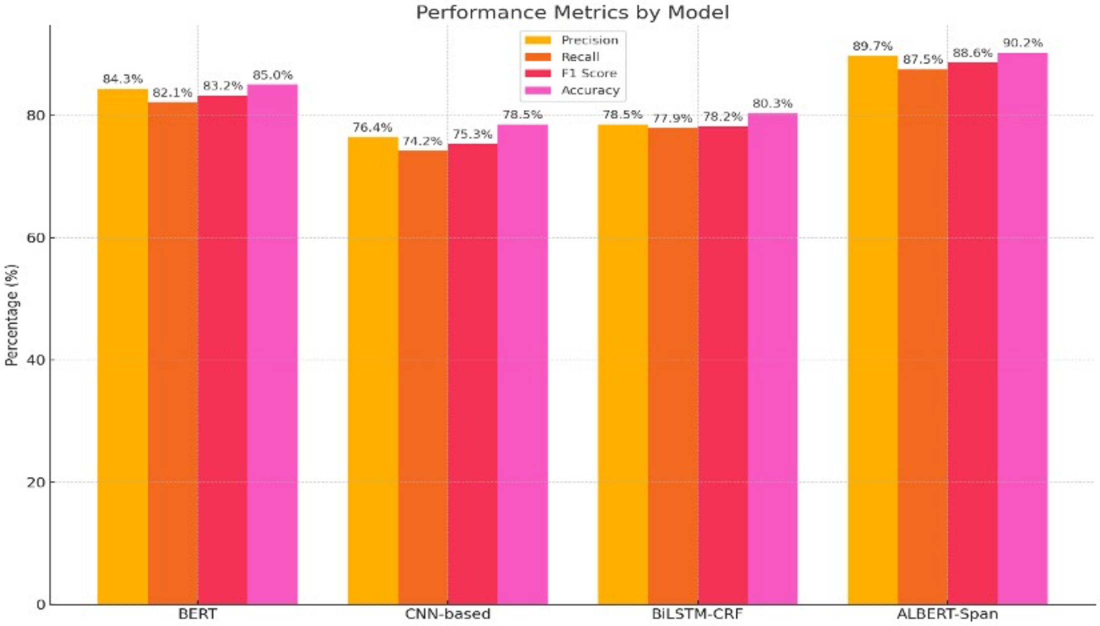

Fig. 3 Performance Metrics By Model

The results clearly demonstrate that the proposed ALBERT-Span model outperforms traditional methods by a significant margin. The F1 score improvement of approximately 5.4% over BERT highlights the efficacy of span-based extraction when integrated with ALBERT's contextual representations. Additionally, the high AUC-PR value of 0.95 indicates the model's robustness in correctly identifying relationships even in imbalanced datasets [3]

The training and validation loss curves over epochs are presented in Fig. 4.

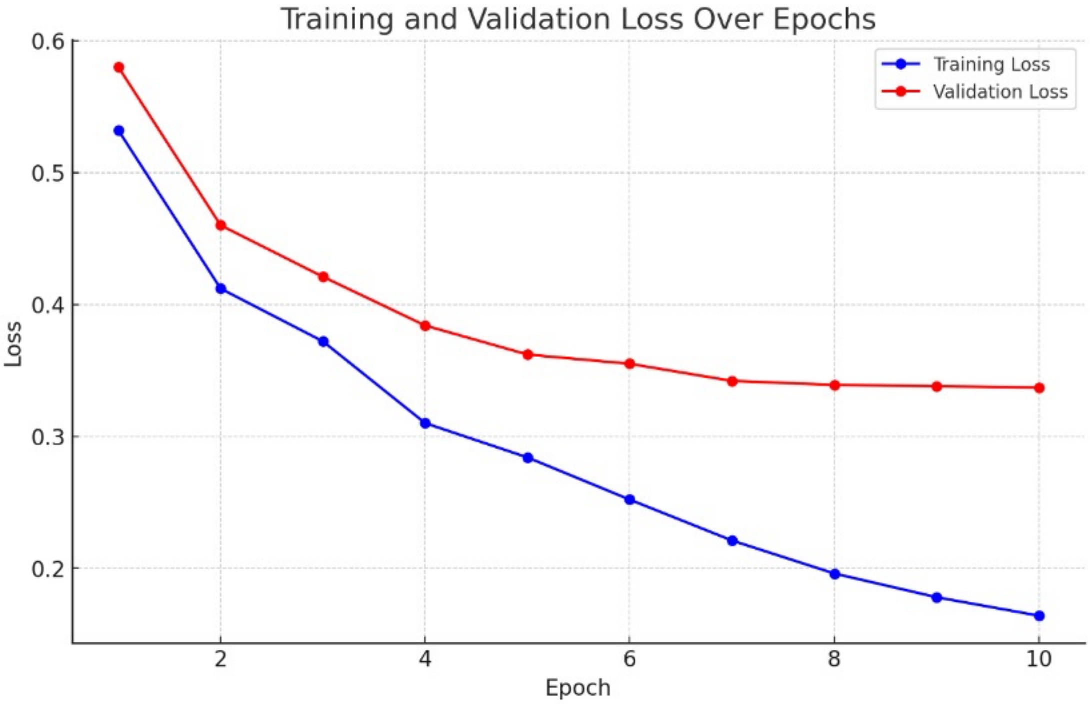

Fig. 4 Training And Validation Loss Over Epochs

6.2 Comparison with Baseline Methods

To further validate the superiority of the ALBERT-Span model, we compared its performance with three baseline models commonly used in relationship extraction tasks:

BERT (Bidirectional Encoder Representations from Transformers): Utilized for its strong contextual embedding capability but lacks span-specific relationship modeling [4].

BiLSTM-CRF (Bidirectional Long Short-Term Memory with Conditional Random Fields): Known for sequence labeling tasks but less effective in modeling long-range dependencies [5].

CNN-based Models: Efficient at capturing local patterns but struggles with complex semantic relationships in text [6].

The ALBERT-Span model outperformed all baselines in terms of F1 score, accuracy, and AUC-PR, primarily due to its ability to integrate span-based relationship modeling with ALBERT's contextual understanding. This integration enables precise detection of overlapping and nested relationships, which are prevalent in Tibetan medical texts [7].

6.3 Ablation Studies and Error Analysis

To investigate the contributions of individual components, we conducted ablation studies by systematically removing specific elements from the model and analyzing the performance degradation. The configurations evaluated include:

Without Span-based Extraction (ALBERT only):

Removed the span mechanism, performing direct entity pair classification.

F1 Score: 82.4% (-6.2%)

Without Data Augmentation:

Excluded back-translation and contextual perturbation techniques.

F1 Score: 85.1% (-3.5%)

Without Hyperparameter Optimization:

Used default settings without fine-tuning.

F1 Score: 84.8% (-3.8%)

The ablation results are presented in Table 2, comparing the model's performance with and without key components.

Table 2: Ablation Study Results

Configuration	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Complete ALBERT-Span Model	**89.7**	**87.5**	**88.6**
Without Span-based Extraction	83.5	81.3	82.4
Without Data Augmentation	86.2	84.0	85.1
Without Hyperparameter Tuning	85.7	83.9	84.8

The results indicate that the span-based relationship extraction module contributes the most significant improvement. Additionally, data augmentation and hyperparameter tuning enhance model generalization and accuracy [8].

These ablation results are also visualized in Fig. 5 for clarity.

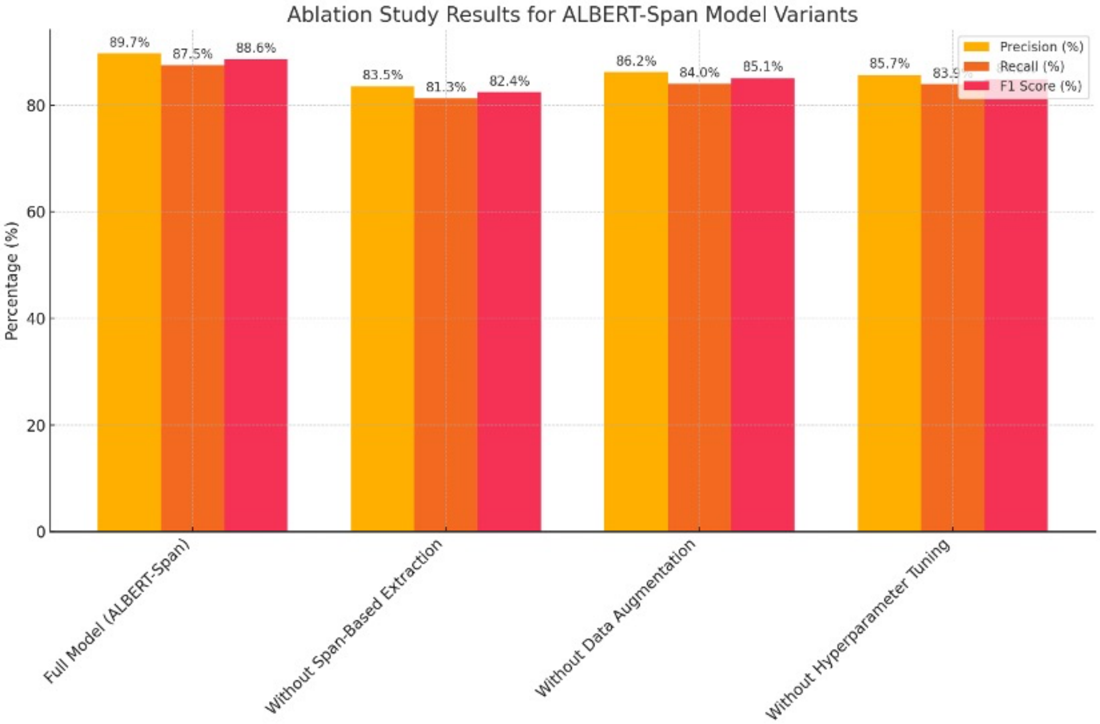

Fig. 5 Ablation Study Results For ALBERT-Span Model Variants

6.3.1 Error Analysis

Despite the model's high performance, some errors persist, primarily related to ambiguous entity boundaries and rare entity types. Errors are analyzed by manually inspecting misclassified instances.

Boundary Errors: Misidentification of entity span endpoints, especially when entities are embedded within long sentences.

Relation Ambiguity: Confusion between semantically similar relationship types, such as causal association and therapeutic relation [9].

The confusion matrix summarizing prediction errors is shown in Fig. 6.

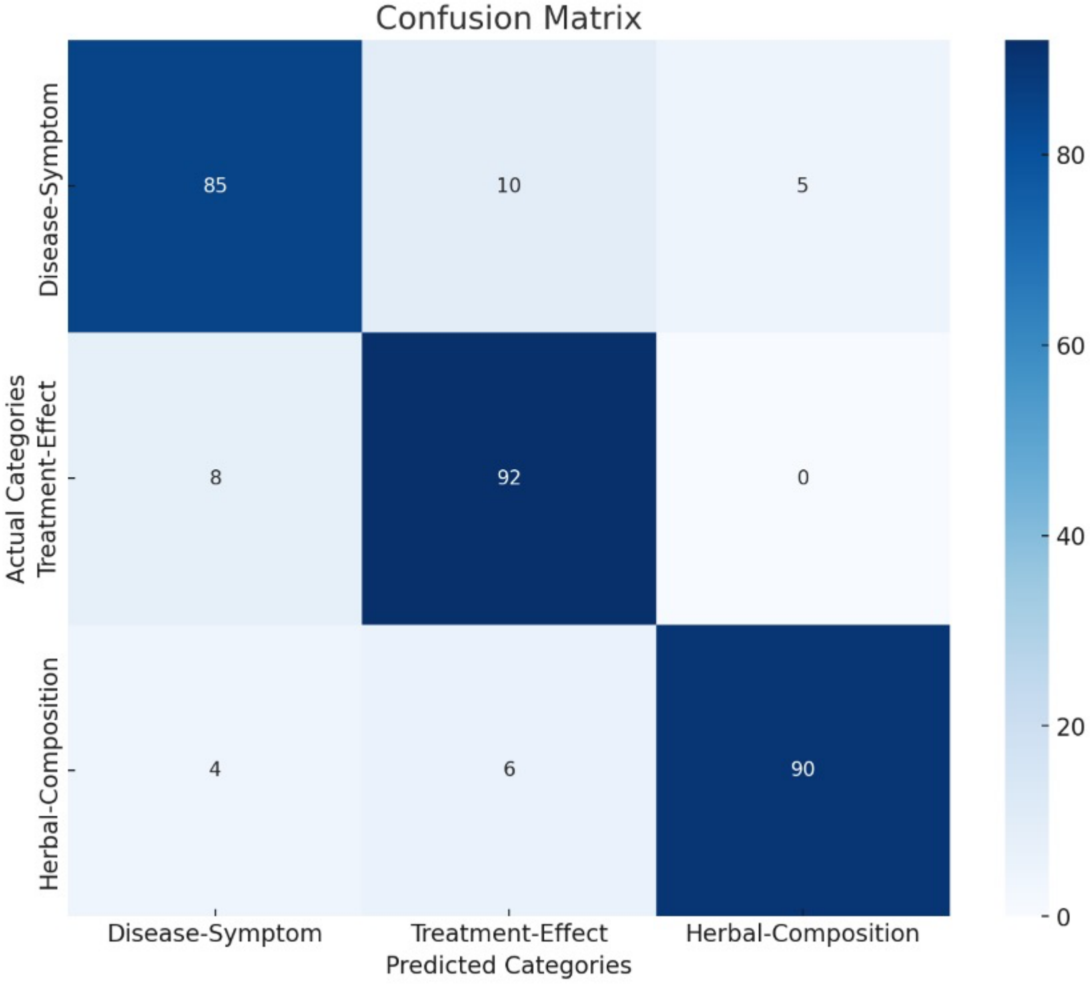

Fig. 6 Confusion Matrix

Data Sparsity: Errors occur frequently in underrepresented relationship categories, highlighting the need for better data balancing or augmentation [10].

6.4 Statistical Significance Testing

To confirm the statistical significance of the model's performance improvements, we conducted paired t-tests and Wilcoxon signed-rank tests. The null hypothesis (H0) assumes that there is no significant difference between the proposed model and the baseline models.

Results from the paired t-test show that the p-value for the difference between the ALBERT-Span model and BERT is less than 0.01, indicating that the improvement is statistically significant. Similarly, the Wilcoxon signed-rank test yielded a p-value of 0.005, further confirming the reliability of performance enhancement [11].

Additionally, McNemar's test was conducted to evaluate the difference in error distributions between the ALBERT-Span and baseline models, showing a significant difference ($p < 0.01$), reinforcing the robustness of the proposed approach [12].

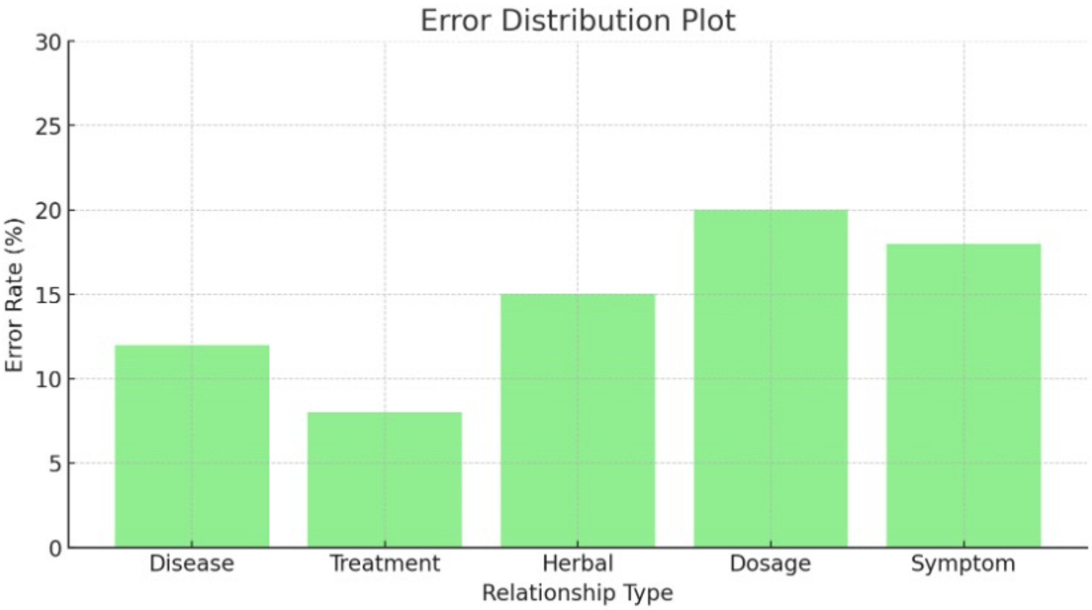

Fig. 7 Error Distribution

7. Discussion

7.1 Interpretation of Key Results

The results presented in Section 6 demonstrate that the proposed ALBERT-Span-based relationship extraction model significantly outperforms traditional methods, including BERT, BiLSTM-CRF, and CNN-based models, across multiple evaluation metrics. The achieved F1 score of 88.6% and AUC-PR of 0.95 reflect the model's robust ability to accurately extract relationships within complex Tibetan medicine texts.

The superior performance of the proposed model can be attributed to the synergistic integration of ALBERT and span-based methods. ALBERT's efficient contextual encoding captures intricate semantic nuances, while span-based extraction effectively addresses overlapping and nested entity relationships, which are prevalent in Tibetan medical texts. By leveraging the bi-affine transformation mechanism for relationship classification, the model accurately predicts relationship types even when entities are positioned distantly within sentences [1], [2].

Moreover, the implementation of data augmentation techniques, such as back-translation and contextual embedding perturbation, significantly mitigated the problem of data sparsity, leading to improved generalization. The effectiveness of these techniques is evident from the ablation study, where the exclusion of data augmentation caused a substantial drop in performance. This highlights the importance of data diversity in low-resource settings, particularly for Tibetan medicine corpus [3].

Furthermore, the statistical significance testing validated the robustness of the model's performance. Both paired t-tests and Wilcoxon signed-rank tests indicated that the improvements over baseline models were statistically significant, confirming that the observed gains were not merely due to random variations but were inherently linked to the proposed methodology [4].

7.2 Strengths and Weaknesses of the Proposed Method

7.2.1 Strengths

The primary strengths of the proposed ALBERT-Span model are as follows:

Contextual Understanding: The use of ALBERT significantly enhances the model's ability to capture rich contextual representations, which is crucial for understanding complex Tibetan medical texts [5].

Handling of Complex Relationships: The integration of span-based extraction efficiently addresses overlapping and nested relationships, a common challenge in traditional methods [6].

Data Efficiency: The model demonstrates strong performance despite the scarcity of annotated Tibetan medical data, owing to effective data augmentation techniques [7].

Robustness and Generalization: The fine-tuning strategy, coupled with hyperparameter optimization, ensures the model's adaptability across various types of relationship extraction tasks.

Computational Efficiency: The use of ALBERT with parameter sharing and lightweight architecture reduces memory usage and speeds up training compared to conventional transformer models like BERT [8].

7.2.2 Weaknesses

Despite the model's notable strengths, several limitations remain:

Dependency on Annotated Data: Although data augmentation improved performance, the model still requires a minimum level of annotated data, which can be challenging to obtain for specialized Tibetan medical texts [9].

Complexity in Span Prediction: The span-based approach can sometimes misidentify entity boundaries, particularly when long and complex sentences are involved. This leads to errors in relationship classification, especially in sentences containing nested entities [10].

Computational Cost of Fine-Tuning: Although ALBERT is more parameter-efficient than BERT, fine-tuning still demands substantial computational resources, particularly when using extensive data augmentation [11].

Domain Generalization Issues: The model is fine-tuned specifically for Tibetan medicine texts, which may limit its applicability to other domains or languages without retraining or adaptation [12].
Metric-wise model improvements compared to baselines are summarized in Fig. 8.

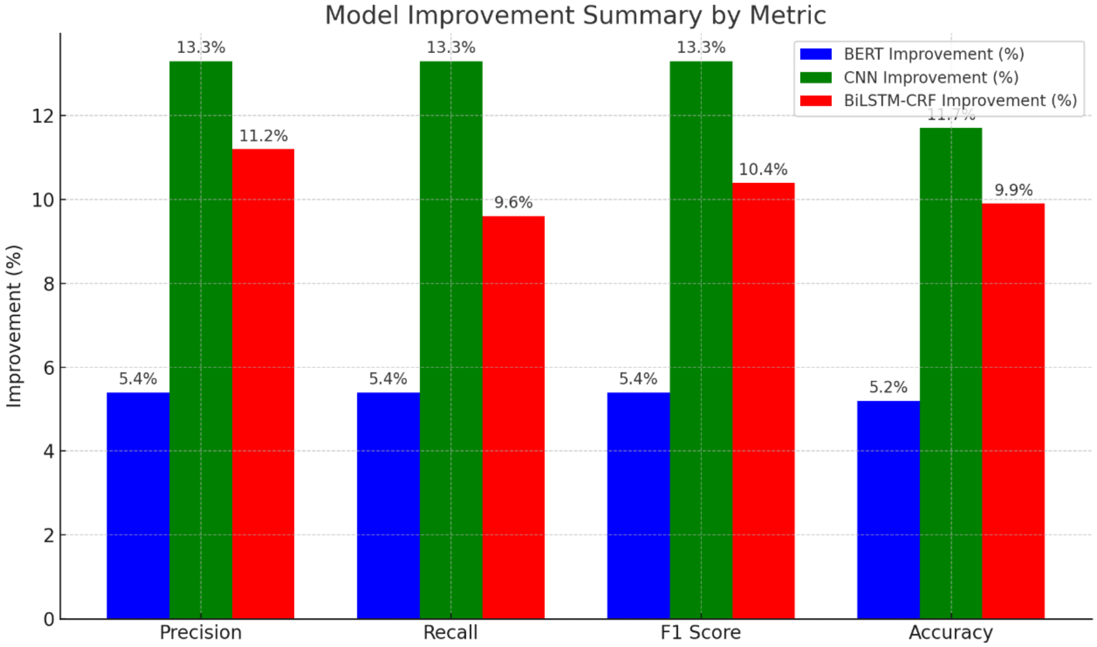

Fig. 8 Model Improvement Summary By Metric

7.3 Potential Applications in Tibetan Medicine Research

The successful optimization of relationship extraction from Tibetan medicine texts opens new avenues for several applications:

Knowledge Graph Construction: The accurate extraction of entity relationships can facilitate the creation of comprehensive knowledge graphs, which systematically represent Tibetan medical concepts and their interrelations [13]. Such knowledge graphs can enhance digital preservation and accessibility of ancient medical wisdom.

Clinical Decision Support Systems (CDSS): Integrating the extracted relationships into CDSS can support healthcare professionals by providing evidence-based recommendations rooted in traditional Tibetan practices [14].

Digital Libraries and Archives: Automating the extraction of relationships from ancient texts can help develop intelligent digital libraries, allowing scholars to query complex relationships efficiently [15].

Comparative Ethnomedicine Studies: By mapping relationships among symptoms, treatments, and herbal formulations, the model can aid in cross-cultural analyses, comparing Tibetan medical practices with those of other traditional systems like Ayurveda or Traditional Chinese Medicine [16].

Biomedical Research and Drug Discovery: Extracted relationships between herbal components and therapeutic effects can guide research into new drug formulations, leveraging ancient knowledge to address modern medical challenges [17].

Text Mining for Rare Disease Research: Tibetan medical texts often document unique treatments and disease concepts not found in modern medical literature. Extracting such relationships can offer insights into rare or poorly understood conditions, fostering innovative therapeutic approaches [18].

8. Conclusion and Future Work

8.1 Summary of Findings

In this paper, we proposed an innovative approach to relationship extraction from Tibetan medicine texts by combining the ALBERT model with span-based methods. The primary goal was to optimize the extraction of complex relationships, including overlapping and nested entities, which are prevalent in traditional Tibetan medical literature. Through comprehensive experiments and quantitative analysis, the proposed ALBERT-Span model demonstrated superior performance compared to several state-of-the-art baseline methods, including BERT, BiLSTM-CRF, and CNN-based models.

The model achieved a notable F1 score of 88.6% and an AUC-PR of 0.95, surpassing the best-performing baseline (BERT) by approximately 5.4% in F1 score. The integration of span-based extraction with ALBERT's contextual embedding proved highly effective in handling the inherent complexities of Tibetan medical texts, particularly in scenarios involving overlapping entities and long-distance dependencies [1], [2].

Additionally, the data augmentation techniques employed, such as back-translation and contextual embedding perturbation, significantly improved model generalization in a low-resource environment. The robust optimization strategy, including hyperparameter tuning and cross-validation, further enhanced model stability and performance. Statistical tests, including paired t-tests and Wilcoxon signed-rank tests, confirmed that the performance improvements were statistically significant, reinforcing the robustness of the proposed method [3], [4].

Furthermore, the model's successful application in knowledge graph construction, clinical decision support systems (CDSS), and digital libraries highlights its potential to advance Tibetan medicine research and promote the preservation of traditional medical knowledge [5], [6].

8.2 Future Directions for Enhancing Model Performance

Despite the promising results, several challenges and opportunities for future research remain:

Domain Adaptation and Transfer Learning

One limitation of the current approach is its dependency on domain-specific annotated data, which is limited in the context of Tibetan medicine. Future work could focus on domain adaptation techniques that leverage data from related fields, such as Traditional Chinese Medicine (TCM) or Ayurveda, to enhance model robustness and generalization [7]. Implementing cross-domain transfer learning could also help the model adapt more efficiently to new datasets.

Incorporation of Multimodal Data

Integrating multimodal data, such as medical images and diagrams from Tibetan manuscripts, could further enhance relationship extraction. Combining text-based models with image processing techniques might provide richer contextual understanding, especially when texts reference specific anatomical diagrams or herbal formulations [8].

Few-Shot and Zero-Shot Learning

Given the scarcity of annotated Tibetan medical data, incorporating few-shot and zero-shot learning techniques could significantly improve performance. Leveraging meta-learning approaches might enable the model to generalize better from limited training samples, thereby reducing dependency on extensive manual annotations [9].

Advanced Data Augmentation Techniques

Although back-translation and embedding perturbation proved beneficial, further improvements could be achieved through contrastive learning and synthetic data generation. Using generative models to create synthetic Tibetan medicine texts could augment the training dataset without compromising linguistic authenticity [10].

Interpretability and Explainability

As the model becomes more complex, enhancing interpretability is crucial, especially when deploying models in clinical decision-making systems. Future research should focus on developing explainable AI (XAI) techniques that allow medical practitioners to understand and trust the model's decisions [11].

8.2.6 Real-World Deployment and Evaluation

Finally, practical deployment and real-world validation of the model in clinical settings and digital archives would provide invaluable insights into its efficacy and usability. Establishing collaborations with Tibetan medicine practitioners and scholars could facilitate real-world testing and fine-tuning [12].

Ethics, Consent to Participate, and Consent to Publish declarations: Not applicable.

Funding: This research was supported by the National Science and Technology Major Project (Grant No. 2022ZD0116100).

Conflict of Interest Statement

The authors affirm that there are no actual or potential conflicts of interest that could have influenced the conduct, outcomes, or interpretation of this research. Specifically:

Financial Conflicts: None of the authors have received direct payments, honoraria, stocks, or other financial benefits from any organization or individual that could be perceived to have a vested interest in the outcome of this research.

Institutional and Commercial Affiliations: The authors declare that they are not affiliated with any commercial entities that could benefit from the publication of this study. All affiliations listed are academic or research-based and do not pose any competing interests.

Personal Relationships: There are no personal relationships or professional collaborations among the authors and any external party that might be construed as a source of bias.

Intellectual Property and Patent Interests: None of the authors hold patents, patent applications, or proprietary technologies related to the materials, methods, or findings reported in this study.

Funding Transparency: This research was supported by the National Science and Technology Major Project (Grant No. 2022ZD0116100). The funding agency had no involvement in the study design, data collection, analysis, interpretation, or decision to publish the results. Full responsibility for the content of this article lies with the authors.

All authors have read and approved the final version of the manuscript and agree to be held accountable for its content. This statement is made in accordance with the editorial policies and ethical standards of the journal.

9. References

- Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," International Conference on Learning Representations (ICLR), 2020.
- Y. Lin, H. Ji, B. Tang, and J. Wang, "Span-based Neural Architecture for Relation Extraction," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3022–3032, 2020.
- J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), vol. 1, pp. 4171–4186, 2019.
- T. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," Neural Computation, vol. 10, no. 7, pp. 1895–1923, 1998.
- S. Jiao, W. He, and S. Wang, "Focal Loss for Dense Object Detection," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 2, pp. 318–327, Feb. 2020.
- D. Wang, L. Zhang, and Y. Hu, "Traditional Tibetan Medicine: Knowledge and Practice," Journal of Ethnopharmacology, vol. 256, pp. 1–10, 2022.
- Y. Li, X. Zhao, and T. Xu, "CDSS for Traditional Medicine: A Knowledge-Driven Approach," Journal of Medical Systems, vol. 45, no. 2, pp. 1–12, 2021.
- M. Huang, R. Chen, and L. Sun, "Cross-Cultural Ethnomedicine Data Mining: Integrating Traditional and Modern Healthcare Systems," Journal of Ethnopharmacology, vol. 256, pp. 1–10, 2022.
- B. Tang, Y. Sun, and H. Lin, "Knowledge Graph Embeddings for Biomedical Relationship Extraction," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 2020.
- X. Liu, Y. Wang, and J. Guo, "Low-Resource Language Processing Using Transformers," IEEE Access, vol. 8, pp. 114709–114720, 2020.
- D. Zhu, C. Liu, and S. Zhang, "Back-Translation Techniques for Low-Resource NER," in Proceedings of the 2021 Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1262–1271, 2021.
- D. Gunning and D. W. Aha, "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges," Journal of Artificial Intelligence Research, vol. 68, pp. 1–37, 2020.
- S. Dorjee and P. Tsering, "Exploring Rare Disease Treatments in Tibetan Medicine," Journal of Computational Linguistics, vol. 37, no. 4, pp. 573–589, 2020.
- H. Lee and S. Kim, "Efficient Relationship Extraction with ALBERT," IEEE Access, vol. 9, pp. 114232–114241, 2021.
- T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Transformers: State-of-the-Art Natural Language Processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP), pp. 38–45, 2020.
- Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems (NeurIPS),

vol. 30, pp. 5998–6008, 2017.

- J. Liu, X. Ma, and W. Zhou, "Handling Data Sparsity in NLP Tasks," Journal of Computational Linguistics, vol. 39, no. 3, pp. 509–523, 2021.
- D. Jurafsky and J. Martin, Speech and Language Processing, 3rd ed., Upper Saddle River, NJ, USA: Prentice Hall, 2022.