2024 International Conference on Asian Language Processing (IALP) | 979-8-3315-4085-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/IALP63756.2024.10661146

Review of Research on Mongolian Fixed Phrases Recognition

Xiaolong Sui, Zhonghao Zhang, Na Liu✉, Guiping Liu, Yatu Ji, Qing-Dao-Er-Ji Ren, Nier Wu, Min Lu
School of Information Engineering
Inner Mongolia University of Technology
Hohhot 010051, P. R. China
csnaliu@imut.edu.cn

Abstract—**Mongolian fixed phrase recognition is one of the most fundamental tasks in Mongolian natural language processing, and its main purpose is to identify the fixed phrase boundaries and types with specific meanings in Mongolian. We introduce the concept, classification, and evaluation of Mongolian fixed phrases at first. Secondly, according to the research process of Mongolian fixed phrase recognition, methods are classified into two categories: dictionary-based and rule-based methods in the past, which are summarized. A detailed comparative analysis reveals the effectiveness and limitations of these methods in practical applications. Then, due to the lack of literature and research, we implement three Mongolian fixed phrase recognition models based on sequences labeling, and briefly introduce the development of Mongolian named entity recognition based on deep learning. Finally, the research trends of Mongolian fixed phrase recognition are discussed to provide some reference for the proposal of new methods and future research directions. Due to the lack of reference studies, we implement a sequences labeling-based model using the available corpus and conduct a discussion and analysis.**

Index Terms—**natural language processing, Mongolian fixed phrase recognition, low-resource language**

I. Introduction

Mongolian fixed phrase recognition is one of the most fundamental tasks in Mongolian natural language processing, and its main content is to identify the fixed phrase boundaries and types with specific meanings in Mongolian. However, Mongolian is a low-resource language that faces challenges such as rich morphology, vague boundaries of fixed phrases, uneven distribution of categories, and scarcity of fixed phrase corpora, resulting in slow advancements in the recognition research of Mongolian fixed phrases. First, this paper introduces the concept of Mongolian fixed phrases, their classification, and their evaluation index. Secondly, according to the Mongolian fixed phrase recognition research process, methods are classified into two categories: dictionary-based methods and rule-based methods in the past, which are summarized. A detailed comparative analysis reveals the effectiveness and limitations of these methods in practical applications. Then, due to the lack of literature and research, we implement three Mongolian fixed phrase recognition models based on sequences labeling, and briefly introduce the development of Mongolian named entity recognition based on deep learning. Finally, the research trends of Mongolian fixed phrase recognition are discussed to provide some reference for the proposal of new methods and future research directions.

However, rich morphology, vague boundaries of fixed phrases, uneven distribution of categories, and scarcity of fixed phrase corpora present challenges for their recognition. Consequently, research on the recognition of Mongolian fixed phrases holds significant importance. This study aims to comprehensively review existing research on Mongolian fixed phrase recognition, analyze the strengths and limitations of various recognition methods, and propose potential directions and trends for future research based on these analyses.

II. Definition and Classification of Fixed Phrases in the Mongolian Language

A. *Definition of Mongolian Fixed Phrases*

The definition of Mongolian fixed phrases [1], [2] includes the following aspects of content,

(1) Fixed phrases are essentially composed of two or more words. The components can be content words, function words, or affixes, but one of the components must be a content word.

(2) Fixed phrases are closely combined in terms of structural form or semantics, generally fixed into a grammatical unit, like a single word. Some fixed phrases are semantically fixed, while others are fixed in form.

(3) As a lexical unit, fixed phrases represent a general concept.

(4) Fixed phrases serve as a sentence element or an auxiliary element within a sentence.

B. *Classification of fixed phrases in Mongolian.*

Fixed phrases in Mongolian are classified into five major categories: compound words, idioms, phrases, fixed expressions, and noun terms [3], [4]. The categories and examples of Mongolian fixed phrases are shown in Table I.

In Mongolian fixed phrases, compound words make up over 95% of the total. Therefore, the recognition of compound words in Mongolian is particularly crucial.

Sui and Zhang contribute equally to this work

TABLE I
MONGOLIAN FIXED PHRASE CATEGORIES AND EXAMPLES

	Categories and Examples	
Mongolian fixed phrases	compound words	ᠠᠰᠠᠭᠤᠳᠠᠯ ᠦᠭᠡᠢ (no problem)
	idioms	[illegible] (good nature)
	phrases	[illegible] (by your will)
	fixed expressions	[illegible] (in terms of...)
	noun terms	[illegible] (lantern slide)

C. The definition and classification of compound words

The study of compound words in Mongolian has a long history, but there is no unified definition. The author has selected the following five classic concepts for reference and learning.

The "Dictionary of Language and Linguistics" defines compound words as "new words formed by combining two or more elements." The most common examples are compound nouns formed by combining two nouns, such as "pencil box" and "handiwork" [5].

In "Modern Mongolian," compound words are "fixed phrases consisting of two or more content words that express noun or verb concepts [6]".

Qing Gertai, in "Modern Mongolian Grammar," defines Mongolian compound words as "lexical units composed of two or more words expressing a single concept [7]".

Deqinggeletu defines compound words as "fixed phrases consisting primarily of two content words (sometimes including auxiliary words) that form a grammatical and lexical unit [8]".

The "Encyclopedia of Mongolian Studies: Language and Script" [9] defines a compound word as "a lexical unit formed by closely combining two or more words to express a fixed meaning." These five classic concepts have largely similar definitions, first and foremost agreeing that "Mongolian compound words are composed of words. "However, there are differences or insufficient expressions regarding the number of component words, the substantive or functional nature of the component words, and the structural fixation. Nevertheless, the combination of substantive and functional elements pertains to usage issues, which this study will not explore in depth.

Compound nouns: These compound words refer to people, things, or phenomena, such as "ᠡᠬᠡ ᠣᠷᠣᠨ" (homeland), "[illegible]" (mountain pass), "[illegible]" (oral stunts), "[illegible]" (dining-table),and "[illegible]" (harmonica). They can be further divided into tangible and intangible nouns, as well as countable and uncountable nouns. They constitute the highest proportion in fixed phrases, accounting for over 80%. Therefore, the focus of Mongolian fixed phrase recognition lies in the identification of compound words.

Compound adjectives: Used to describe qualities, shapes, and states, for example, "[illegible]" (brag), "[illegible]" (frank), "[illegible]" (stubborn), "[illegible]" (beautiful). According to their meaning, form, and function, compound adjectives can be subdivided into qualitative adjectives, relational adjectives, and color adjectives.

Compound verbs: They express actions, behaviors, states, and changes, mainly composed of lexical verbs, while auxiliary verbs, causative verbs, and linking verbs do not participate in compounds. For example, "[illegible]" (is studying), "[illegible]" (is resting).

Compound temporal and locative nouns: They express the time or place of action, such as "[illegible]" (before and after), "[illegible]" (left and right), "[illegible]" (inside and outside), "[illegible]" (up and down). Compound temporal and locative nouns possess characteristics of both content words and function words.

Compound pronouns: They replace people, things, or their qualities, characteristics, quantities, times, or places, such as "[illegible] " (back and forth), "[illegible]" (who). Compound pronouns do not directly refer to specific things but have a referential nature, thus exhibiting abstract and generalized characteristics, reflecting the attributes of the word class they replace in context.

Compound adverbs: They modify the tense, form, or state of actions, such as "[illegible]" (long ago), "[illegible]" (uneven), "[illegible]" (eternally), "[illegible]" (hastily). Compound adverbs do not refer to specific things but modify verbs or adjectives, their number is relatively small.

III. OVERVIEW OF MONGOLIAN FIXED PHRASE RECOGNITION

A. Mongolian Fixed Phrase Corpora

Due to the absence of publicly available Mongolian fixed phrases corpora, the following outlines the relevant corpora mentioned in the research literature.

Si Laogelao [10], in the paper "Design and Implementation of Mongolian Fixed Phrase Recognition Algorithm," constructed a Mongolian fixed phrase recognizer based on the "Mongolian Fixed Phrase Grammatical Information Dictionary" [2] and the "Inflectional Suffix Component Dictionary." Six volumes of high school Mongolian language textbooks, selected from the "Modern Mongolian Corpus" built by the School of Mongolian Studies at Inner Mongolia University, were used as test corpora. Researchers manually annotated the test corpora, marking 275,008 Mongolian fixed phrases.

Wang et al. [11] formulated the Mongolian-named entity annotation guidelines, establishing the scope, types, and principles of annotation during the research on Mongolian named entity recognition. The Brat tool was used to build the annotation platform. Using this platform, 33,292 sentences were annotated, resulting in 59,562 Mongolian named entities, with place names accounting for 47.62%, institution names for 31.64%, and personal names for 20.74%.

Liu et al. [12] provided a Mongolian syntactic analysis corpus containing 5,000 Mongolian sentences, including 7,529 fixed phrases. We named this corpus MFPC, and its details are shown in Table II.

TABLE II
MFPC

category	number
compound noun	7140
compound adjectives	289
compound pronoun	0
compound time-place words	4
compound verb	86
compound adverb	18
adjective idioms	0
verbal idiom	1
nominal idiom	0
verbal idiom	0
phrase	0
fixed word-combination	0
terminology	1

B. Evaluation Metrics for Mongolian Phrase Recognition

Precision: This accuracy metric measures the system's ability to correctly recognize Mongolian phrases. It represents the proportion of actual positive instances among all instances predicted as positive by the model. The formula for calculating precision is:

$$\text{Precision} = \frac{\text{Correct }_{\text{num}}}{\text{Recognize }_{\text{num}}} \times 100\% \tag{1}$$

Correct: Number of correctly identified fixed phrases. Recognize The total number of recognized fixed phrases.

Recall: The recall metric measures the proportion of actual positive instances the model correctly identifies. The formula for the recall is:

$$\text{Recall} = \frac{\text{Correct }_{\text{num}}}{\text{Actual }_{\text{num}}} \times 100\% \tag{2}$$

$F1-score$: The $F1-score$ is the harmonic mean of precision and recall, used to evaluate the model's performance comprehensively. When both precision and recall are high, the F-score will also be high. The formula for the $F1-score$ is:

$$\text{F1} - \text{score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\% \tag{3}$$

When evaluating model performance, these metrics are typically considered comprehensively. For example, if a model shows high precision and recall in compound word recognition and has a high $F1-score$, we can conclude that the model performs well in recognizing compound words. Additionally, we need to be aware of support differences across categories to ensure the fairness and accuracy of the evaluation results.

IV. Identification Methods for Mongolian Fixed Phrases

A. Dictionary-Based Recognition Method

The dictionary-based recognition method is one of the earliest techniques to identify fixed phrases. However, this method requires a substantial dataset to be effective. In the face of the lack of traditional Mongolian dictionaries, researchers [13] proposed an innovative strategy for dictionary construction. Recognizing and statistically analyzing the frequency of terms in a vast corpus of ancient Mongolian texts, they meticulously selected high-frequency words. They incorporated them into a newly constructed dictionary. This strategy laid a solid foundation for subsequent error detection and correction tasks and provided valuable resources for the in-depth understanding and study of ancient Mongolian literature. Dictionary-based methods can effectively perform error detection tasks by utilizing this constructed dictionary. The system first matches the recognition results with the entries in the dictionary, allowing for precise identification and marking of non-word errors (i.e., character sequences that do not form legitimate words). The researchers introduced a weighted edit distance model and a noisy channel model for error correction to enhance recognition accuracy. These advanced models evaluate the similarity between recognition results and dictionary entries, intelligently recommending or replacing incorrectly recognized words, thereby significantly improving recognition accuracy. Experimental results demonstrate that dictionary-based error detection and correction methods have significantly increased word recognition accuracy, from 80.86% to 84.79%. This improvement proves the effectiveness of dictionary-based methods and brings new hope to the digitization and preservation of ancient Mongolian literature. Despite the significant achievements in improving recognition accuracy, dictionary-based methods still have limitations in detecting real-world errors. Additionally, the completeness and quality of the dictionary directly affect the effectiveness of error detection and correction. Therefore, it is imperative to establish large datasets to support the recognition of Mongolian fixed phrases. Future research should continue to expand and refine dictionary resources and explore more advanced error detection and correction algorithms. This will help improve the system's recognition accuracy and robustness, further advancing the digitization of ancient Mongolian texts.

B. Rule-based Method

People in the early days constructed a fixed phrase recognizer using the prioritized state machine automaton method based on the "Mongolian Fixed Phrase Grammar Information Dictionary" and the "Inflected Additional Component Dictionary". This recognizer consists of the fixed phrase automaton (FPA) and the formation suffix automaton (FSA).

In construction, both FPA and FSA are the same, both in the form of a 9-tuple as follows:

$$FPA = FSA = \left(S, \sum, \delta, \text{ suffix, recall, stack}, S_0, U, F\right) \tag{4}$$

where S represents a finite set of states, represents a finite set of input characters, is the state transition function, and suffix is the state transition function from the fixed phrase root automaton to the additional component automaton. The recall function is used for backtracking to the root, and the stack is used to store roots. S_0 is the initial state, U is the set of root states, and F is the set of final states [10]. For the recognition of fixed Mongolian phrases, a Mongolian fixed phrase recognizer can also be constructed. The first step of the algorithm is to construct a finite state automaton (FPA and FSA), which involves defining the state set, input character set, and state transition function. Next, the text to be analyzed is input into the FPA for processing. During the recognition process, when encountering a root state, it is pushed onto the root stack and transferred according to the state transition function. If a state transition cannot be performed, the current character is checked to see if it is a space, if it is, fixed phrase annotation is performed, otherwise, root backtracking is executed. When a state transition cannot be performed and the root stack is not empty, the state is popped from the top of the stack and the state transition is performed through the FSA, after which recognition starts again from the root state. Once a fixed phrase is recognized, appropriate annotation or other post-processing is performed. The experimental results are shown in Table III.

TABLE III
PERFORMANCES OF RULE-BASED METHOD

Correct $_{\text{num}}$	Actual $_{\text{num}}$	R (%)
20623	18507	89.7

The experimental results indicate that the rule-based model achieves 89.7% in Recall, demonstrating high efficiency. Additionally, collecting fixed phrases from large-scale text corpora through manual or semi-automated methods to expand the ”Mongolian Fixed Phrase Grammar Information Dictionary” can further improve the algorithm’s recall rate. Furthermore, due to the morphological variations of fixed phrases in Mongolian, the algorithm’s adaptability is also considered extendable to other dictionary-based Mongolian text processing tasks.

Despite its advantages in efficiency, flexibility, and accuracy, the algorithm also faces certain challenges and limitations. Its construction and state transition complexities require relevant expertise and experience for effective application. Additionally, the algorithm’s performance heavily depends on the quality and completeness of the dictionary. In handling morphologically diverse fixed phrases, the root backtracking mechanism may negatively impact the algorithm’s efficiency.

C. Recognition of Mongolian Fixed Phrases Based on Sequentials Labeling

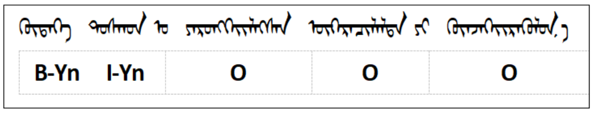
The development of computer technology has revealed the limitations of traditional dictionary-based and rule-based recognition methods, which are often ineffective for entities beyond predefined rules and are costly to maintain. The emergence of big data has prompted a shift towards statistical models for training and prediction, which excel in handling complex linguistic phenomena without being constrained by rules.

Statistical models, such as Conditional Random Fields (CRF) and Hidden Markov Models (HMM), automatically learn from large annotated corpora to accomplish recognition tasks. CRF, a discriminative model, excels in sequences labeling tasks, such as Mongolian named entity recognition, by modeling the joint probability of label sequences using contextual features. In contrast, HMM, a generative model, infers the most likely state sequence from observed data using transition and observation probabilities.

However, the performance of statistical models can be influenced by the quality of feature engineering. To enhance recognition performance, researchers are exploring advanced methods for automatic feature learning to capture richer representations.

In response to the challenges faced by the rule-based methods mentioned above, the advancement of deep learning technology has offered more feasible solutions for the identification of fixed phrases in Mongolian. This paper, based on the MFPC [12] corpus, has implemented a method for recognizing Mongolian fixed phrases using sequential labeling. This includes the recognition of Mongolian fixed phrases based on Hidden Markov Models (HMM), Conditional Random Fields (CRF), and a combination of Bidirectional Long Short-Term Memory (BiLSTM) and CRF networks. We divided the MFPC into a training set and a test set. The training set consists of 3,500 sentences (70%), while the test set comprises 1,500 sentences (30%). The data was trained and tested in accordance with these proportions. The specific experimental processes and methods are as follows.

1) Data Preprocessing:
In the Mongolian corpus, the data is annotated using the BIO scheme, where “B” stands for “Begin” indicating the start of a fixed phrase, “I” stands for “Inside” indicating the interior of a fixed phrase, and “O” stands for “Outside,” indicating that the token does not belong to any fixed phrase. Taking a nominal compound within a fixed phrase as an example,which is annotated based on the relationships between words, and the result is as follows:



B-Yn	I-Yn	O	O	O

Fig. 1. Sequences Labeling Sample

2) *HMM Based Model:*

In the Hidden Markov Model (HMM), the state set Q is computed from the processed corpus, denoted as λ. The state set Q represents the categories of Mongolian fixed phrases. During the recognition process, a segment of Mongolian text to be identified is treated as the observed sequence O. The Viterbi algorithm is then used to solve for the state sequence I that maximizes $P(I|O)$ given λ and 0, thus obtaining the results of fixed phrase annotation.

3) *CRF Based Model :*

For Conditional Random Fields (CRF), after the data is prepared, the annotated dataset is used to train the CRF model. Once training is complete, the model parameters are obtained, and the model is saved. The saved model is then used to identify new input text. The recognition process is similar to the HMM model, but the CRF model considers more factors at output, not just the previous state, but also all surrounding states, especially the dependencies and contextual information between characters. Assume X and Y represent the input in Mongolian and the output result respectively. Given X, the conditional probability formula for each possible value in the sequence Y is as follows:

$$P(Y \mid X) = \frac{1}{Z(X)} \exp(\sum_{i=1}^{n}\sum_{k=1}^{n} \lambda_k f_k(y_{i-1}, y_i, x, I)) \qquad (5)$$

where Z is the normalization factor, the formula is as follows:

$$Z(X) = \sum_{y} \exp(\sum_{i=1}^{n}\sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, x, i) \qquad (6)$$

4) *BiLSTM+CRF Based Model:*

In the BiLSTM+CRF based model, BiLSTM assists the model understand the relationship between each word in the text, while CRF is used to find out how different parts of the text relate to each other. Specifically, two-way LSTM network is used to process vectors, to capture the text information before and after the word, and finally decode the LSTM output through the CRF layer to obtain the final annotation result.

5) *Results and Analysis:*

The results based on the sequences labeling method are shown in Table IV.

TABLE IV
PERFORMANCES OF SEQUENCES LABELING MODELS

	P(%)	R(%)	$F1 - score(\%)$
HMM	91.3	91.7	91.4
CRF	92.2	92.6	92.3
BiLSTM+CRF	93.4	94.4	93.8

- By comparing the results, all three models demonstrate strong recognition capabilities and effectiveness. Especially the BiLSTM+CRF model achieves the best performance of Mongolian fixed phrase recognition with 93.4% Precision, 94.4% Recall and 93.8% $F1 - score$.
- As a powerful generative model, HMM achieves excellent performance compared to traditional dictionary-based and rule-based methods, without using large-scale dictionaries. However, HMM can not consider the label dependencies of the entire sequence, addressing the shortcomings of HMM and further enhancing recognition accuracy. CRF performed the $F1 - score$ reaches 92.3%, which is a 0.9% improvement compared to the HMM based model. Nevertheless, CRF faces resource wastage issues when recognizing simple words.
- In contrast, the BiLSTM+CRF model combines the advantages of deep learning and statistical machine learning. BiLSTM captures long-term dependencies in sequences and automatically learns effective feature representations. The CRF layer further considers the dependencies between labels, improving prediction accuracy and addressing the issues encountered by HMM and CRF in the recognition process. Therefore, the BiLSTM+CRF model typically achieves better performance.

V. MONGOLIAN NAMED ENTITY RECOGNITION

Although deep learning has yielded numerous outstanding results in the field of natural language processing, the recognition of Mongolian fixed phrases has advanced at a slow pace due to the language's low-resource status and other unique characteristics. Fortunately, researchers have extensively explored Mongolian named entities, a vital component of Mongolian compound noun phrases, using deep learning techniques. This section will provide an overview of these research endeavors and their contributions to the field.

Wang et al. [14]construct the first Mongolian NER corpus and investigate three morphological processing methods and features, including syllable features, lexical features, contextual features, morphological features, and semantic features. Experimental results show that segmenting each suffix into a separate token yields better results. The $F-measure$ reaches 84.65% when combined all features including handcraft features and word cluster features. Following this work, Cheng et al. [15] construct a corpus for Mongolian tourism NER-MTNER. The researchers trained in-domain BERT representations with unannotated Mongolian corpus, and trained a NER model based on the BERT tagging model. The MTNER size reaches 1,600 sentences and 18 entity types, and the $F1-score$ of the proposed model achieves an overall 82.09%.

Xiong et al. [16] find that existing pre-trained language models cannot fully capture the semantics and syntactic roles of context words in labeled data. For the integration of external information, existing methods simply concatenate pre-trained language model embeddings with internal information, failing to effectively balance the differences between them. So [16] proposes LM-ATT (Language Model with Attention) model for Mongolian NER task. The LM-ATT uses attention mechanisms to dynamically balance pre-trained language model embeddings and internal information. Achieves significant performance improvements on Mongolian NER tasks.

Mongolian, as a low-resource language, faces significant challenges in model training due to insufficient labeled data and rich morphological variations, which lead to data sparsity.

Traditional NER models rely on manually crafted features and high-quality corpora, making them ineffective for out-of-vocabulary (OOV) words.

Literature [17] combines Mongolian linguistic knowledge with cross-lingual knowledge to address data scarcity and exploit Mongolian-specific features. By decomposing words into roots and affixes, researchers capture finer-grained semantic information and reduce vocabulary size. We also transfer semantic information from Chinese to Mongolian using parallel corpora, enhancing Mongolian word embeddings. Additionally, we generate pseudo-labeled data through cross-lingual annotation projection, expanding training data and improving model performance. The Multi-Knowledge Enhanced NER (MKE-NER) model achieves an $F1-score$ of 94.04%, significantly outperforming other baseline models and demonstrating its effectiveness in low-resource scenarios.

VI. Prospects and Summary

A. Prospects

In line with recent advancements in the field of natural language processing, the recognition of Mongolian fixed phrases also follows the following trends (1) the application of pre-trained models, (2) cross-lingual processing, and (3) the optimization of deep learning models. Firstly, the development of technology in the field of deep learning, particularly pre-trained models that aid in extracting rich information from input text, secondly, as Mongolian is a low-resource language, cross-lingual processing allows for the transfer of knowledge from high-resource languages to low-resource language tasks, which helps to enhance task performance. Lastly, deep learning remains the mainstream approach in natural language processing, and optimizing these models to improve performance is a necessary path to enhancing task performance.

1) Cross-lingual Data Annotation:
In constructing deep learning-based tasks for the recognition of Mongolian fixed phrases, a major challenge is the lack of large and high-quality annotated corpora. The scarcity of these resources limits the training and performance optimization of models, as deep learning models typically require a substantial amount of data to learn complex feature representations. To overcome this obstacle, current researchers are exploring cross-lingual text processing techniques. For instance, [18] achieves knowledge transfer from high-resource languages to low-resource languages by combining translation, bilingual embeddings, and transliteration strategies to enhance cross-lingual named entity recognition. [19] utilizes cross-lingual contextual word embeddings and transfer learning, applying models trained on English datasets to resource-poor languages such as Arabic, Bengali, Danish, etc., for the identification of offensive language, and has achieved results comparable to the best systems in multiple shared tasks. Therefore, future construction of Mongolian fixed phrase corpora can draw on cross-lingual text processing methods for further research.

2) Unsupervised Methods:
To address the issue of data scarcity, current research efforts are attempting to study unsupervised learning methods. These methods leverage a large number of unlabeled sentences and a small number of independent samples for model training, thereby reducing dependence on expensive and time-consuming manually annotated data [20]. Specifically, unsupervised learning strategies can automatically extract features and patterns from unlabeled text while using a small number of samples to guide the learning process, enhancing model performance and generalization capabilities with limited resources. Additionally, this approach helps to uncover potential structures in the data, providing more viable solutions for NLP tasks in low-resource languages. For example, [21] proposes an unsupervised learning method that combines progressive self-training and a discriminator to address the lack of annotated data in aspect term extraction tasks. By using progressive self-training and dividing the unlabeled dataset into multiple subsets based on difficulty and quantity, and then gradually introducing harder samples, the method introduces a discriminator to filter out noise in pseudo-labels, improving the quality of pseudo-labels generated during the self-training process. This work demonstrates that unsupervised methods have strong generalization capabilities and effectiveness in dealing with the scarcity of annotated data, and unsupervised techniques can be employed in future research to address the scarcity of Mongolian fixed phrase corpora.

3) Guidance of Linguistic Knowledge:
Due to the complex morphology, large vocabulary, and numerous loanwords in Mongolian, the scarcity of corpus resources prevents deep neural network models from learning optimal representations. To address these issues, researchers have introduced Mongolian morphological knowledge into Mongolian information processing tasks, enhancing the performance of related tasks. For instance, [22] proposes learning Mongolian morpheme vectors from large-scale corpora to overcome the sparsity of Mongolian word training data, thereby improving the performance of Mongolian named entity recognition tasks, Liu et al. [23] propose a morpheme-based Mongolian prosody modeling method for enhancing the naturalness of Mongolian speech synthesis. These works have studied Mongolian morphological knowledge and proposed solutions suitable for Mongolian, which can be combined with Mongolian morphological knowledge in future research to help address issues in the recognition of Mongolian fixed phrases.

4) Combination of Rules and Neural Networks:
While neural network-based methods are currently the mainstream in natural language processing, the research value of traditional Mongolian fixed phrase recognition methods should not be overlooked. In the OpenIE field, researchers have found that when comparing across models based on factual benchmarks, neural models perform worse than their rule-based counterparts on multiple metrics. While efforts to improve neural network extraction are ongoing, considering the introduction of rule-based methods may yield better results [24]. Currently, how to integrate rule-based models into deep neural network-based recognition models remains an area for in-depth research and exploration.

5) Research on Mongolian Fixed Phrase Recognition Based on Other Deep Learning Methods:

In addition to the mainstream methods of RNNs, CNNs, and Transformers for sequence data annotation using deep learning, other deep learning methods have also been proposed. Literature [25] introduces the ZeroAE model into natural language processing tasks to address challenges encountered in zero-shot classification tasks. The ZeroAE model designs an autoencoder based on a pre-trained language model, combining encoding and decoding methods. This approach promotes complementarity and self-adaptation between the two, representing text as a tensor that not only has size but also direction, which can well maintain the spatial information of the model. Experimental results show that the ZeroAE model has significant advantages in handling unseen domains and a large amount of unlabeled data. By analyzing Table II, it is not difficult to find that the distribution of various Mongolian fixed phrases is not uniform. For example, the corpus contains only 4 compound tense-time words and does not include compound pronouns. When using deep neural networks to establish mathematical models for classification tasks, the sample size of these categories of compound words is too small, leading to an inability to accurately depict the distribution of each category of samples. Therefore, using methods such as the ZeroAE model to address the problem of uneven sample distribution in Mongolian compound word recognition tasks is a worthy research direction.

B. Summary

This paper reviews and summarizes the research on Mongolian fixed phrase recognition. It analyzes the current research status from the concepts, classification, corpora, evaluation criteria, and recognition methods of Mongolian fixed phrases, compares the advantages and disadvantages of research methods, and predicts future development directions. The prospects for Mongolian recognition still have significant potential value.

Acknowledgment

This work was funded by research program of science and technology at Universities of Inner Mongolia Autonomous Region (Grant No. NJZZ21061, NJZZ22251), 2022 Inner Mongolia Talent Support Project(DC2300001440), Inner Mongolia Natural Science Foundation (2024MS06017, 2022MS06013), Science Research Foundation of Inner Mongolia University of Technology(DC2300001262).

References

[1] Nadaamd, “Research on Modern Mongolian Written Compound Words”. Central University for Nationalities, 2019.

[2] De Qinggeltu, “Detailed Explanation of the Grammar Information Dictionary of Modern Mongolian Fixed Phrases”. Hohhot: Inner Mongolia Education Press, 2005.

[3] De Qinggeltu, “Research on Mongolian Fixed Phrases for Information Processing”. Hohhot: Inner Mongolia Education Press, 2001.

[4] De Qinggeltu, “Classification of Mongolian Fixed Phrases for Information Processing”. Journal of Inner Mongolia Normal University (Philosophy and Social Sciences Edition in Mongolian), 2000.

[5] Hartmann, “Dictionary of Linguistics and Language”. Shanghai: Shanghai Dictionary Publishing House, 1981.

[6] Inner Mongolia University, “Modern Mongolian” [M]. Inner Mongolia: Inner Mongolia People’s Publishing House, 2005.

[7] Qinggeltai, “Modern Mongolian Grammar” [M]. Inner Mongolia: Inner Mongolia People’s Publishing House, 1999.

[8] De Qinggeltu, “Description of the Grammatical Attributes of Mongolian Compound Words,” “Journal of Inner Mongolia Normal University,” 2003, Vol. 4.

[9] “Encyclopedia of Mongolian Studies” . Inner Mongolia: Inner Mongolia People’s Publishing House, 2010.

[10] Sa Logolo, “Design and Implementation of the Recognition Algorithm for Mongolian Fixed Phrases”. Journal of Chinese Information Processing, 2017, 31(05): 85-91.

[11] Wang W, F Bao, Gao G. Mongolian Named Entity Recognition with Bidirectional Recurrent Neural Networks. Proceedings of the IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), 2016: 495–500.

[12] Liu N, Su X, Gao G, etc. Morphological Knowledge Guided Mongolian Constituent Parsing. Proceedings of the Neural Information Processing, 2019: 363-375

[13] Su Dongxiang, “Research on the Recognition of Ancient Mongolian Manuscripts Based on Deep Learning and Knowledge Strategies” [D]. Inner Mongolia University, 2016.

[14] Wang, Weihua, Feilong Bao, and Guanglai Gao. ”Mongolian named entity recognition system with rich features.”Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 2016.

[15] Cheng X, Wang W, Bao F, et al. ”MTNER: A Corpus for Mongolian Tourism Named Entity Recognition”. 2020. DOI:10.1007/978-981-33-6162-1-2.

[16] Xiong, Yuzhu, and Minghua Nuo. ”Attention-based blstm-crf architecture for mongolian named entity recognition.”Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation. 2018.

[17] Zhang, Songming, et al. ”Exploiting Morpheme and Cross-lingual Knowledge to Enhance Mongolian Named Entity Recognition.”Transactions on Asian and Low-Resource Language Information Processing 21.5 (2022): 1-19.

[18] Xiaolei Huang, Jonathan May, and Nanyun Peng. 2019. What Matters for Neural Cross-Lingual Named Entity Recognition: An Empirical Analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6395–6401, Hong Kong, China. Association for Computational Linguistics.

[19] Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual Offensive Language Identification for Low-resource Languages. ACM Trans. Asian Low-Resour. Lang. Inf. Process. 21, 1, Article 4 (January 2022), 13 pages. https://doi.org/10.1145/3457610

[20] Andrea Iovine, Anjie Fang, Besnik Fetahu, Oleg Rokhlenko, and Shervin Malmasi. CycleNER: An Unsupervised Training Approach for Named Entity Recognition. In Proceedings of the ACM Web Conference 2022 (WWW ’22). (2022)

[21] Qianlong Wang, Zhiyuan Wen, Qin Zhao, Min Yang, and Ruifeng Xu. Progressive Self-Training with Discriminator for Aspect Term Extraction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. (2021)

[22] Wang W, Bao F, Gao G. Learning morpheme representation for mongolian named entity recognition. Neural Processing Letters, 2019, 50(2): 2647–2664.

[23] Rui L , Bao F , Gao G , etc. Mongolian text-to-speech system based on deep neural network. Proceedings of the Man-Machine Speech Communication, 2017: 99–108.

[24] Gashteovski, Kiril, Mingying Yu, Bhushan Kotnis, Caroline V. Lawrence, Goran Glavas and Mathias Niepert. “BenchIE: Open Information Extraction Evaluation Based on Facts, Not Tokens.” ArXiv abs/2109.06850 (2021)

[25] Kaihao Guo, Hang Yu, Cong Liao, Jianguo Li, and Haipeng Zhang.ZeroAE: Pre-trained Language Model based Autoencoder for Transductive Zero-shot Text Classification. In Findings of the Association for Computational Linguistics: ACL 2023.