CiteSpace-based Research Progress and Trend Analysis of Mongolian-Chinese Machine Translation

Dong Xi[1] Limin Li[2]

[1,2] *Xi'an University of Science and Technology, Xi'an, China*
[2]*Corresponding author. Email: 15565230134@163.com*

ABSTRACT
Machine translation for low-resource languages is a key focus in translation studies. In this paper, CiteSpace was employed to perform a bibliometric analysis of 109 core articles on Mongolian-Chinese machine translation retrieved from CNKI between 2002 and 2024. This analysis systematically maps the latest progress and research trends in this field. The results show that research on Mongolian-Chinese machine translation in China has evolved through three distinct phases: initial exploration, fluctuating growth, and stable development. Although the research on Mongolian-Chinese machine translation has achieved remarkable results, its translation performance still needs to be improved compared with high-resource language models. Current research in this field focuses on Neural Machine Translation. Researchers have primarily focused on methods such as adversarial learning, dual learning, transfer learning, data augmentation, and pre-training to continuously train, iterate, and optimize translation systems, with the goal of enhancing translation performance.

Keywords*: Low-resource language, Machine translation, Mongolian-Chinese machine translation, Bibliometric analysis, Knowledge graph, Low-resource language pairs.*

1. INTRODUCTION

In recent years, technological innovations in artificial intelligence for natural language processing have significantly advanced machine translation technology. The field has evolved from early Rule-Based Machine Translation (RBMT) to Statistical Machine Translation (SMT) and, more recently, to Neural Machine Translation (NMT). Continuous advancements in related technologies have driven significant improvements in both translation efficiency and quality.

Existing bibliometric studies in machine translation predominantly concentrate on summarizing research in a general sense. However, systematic literature reviews specifically addressing Mongolian-Chinese—a low-resource language pair—remain scarce compared to studies on high-resource pairs such as English-Chinese or English-German. For instance, Li Xiang et al. (2024) employed bibliometric methods to analyze 2,295 machine translation-related articles published in the Web of Science (1958-2022), identifying four major research themes: example-based machine translation, statistical machine translation, neural machine translation, and machine translation applications, while outlining three development stages of research frontiers in machine translation. However, the study did not cover the status quo of machine translation for low-resource language pairs, remaining confined to general analyses of the field.[1] Additionally, Fu Linling et al. (2023) conducted a bibliometric analysis of machine translation (MT) literature from CNKI (1992-2022). The study shows steady growth in domestic MT research, with a significant increase after 2016. Key research trends focus on addressing low-quality translations in low-resource language pairs, using methods like big data, data augmentation, transfer learning, and reverse translation, leveraging high-resource parallel corpora.[2] Although touching upon the topic of machine translation for low-resource language pairs, it merely summarized methods proposed by certain scholars to enhance low-resource translation quality, which still has limitations in scope. A recent statistical analysis by Zhang Renran (2024) et al. on intelligent translation literature from China National Knowledge

Infrastructure (CNKI) spanning 1959-2022 demonstrates that the global publication volume in intelligent translation research has maintained an upward trend, accompanied by substantial improvements in translation quality. However, the study also pointed out existing problems: "English-Chinese translation remains the primary focus, with comparatively limited attention given to minor languages and domestic ethnic minority languages."[3]

Given this context, the present study employs CiteSpace to conduct a systematic visual analysis of 109 core articles on Mongolian-Chinese machine translation indexed in CNKI between 2002 and 2024. Through a combined quantitative and qualitative approach, it analyzes four critical dimensions: evolutionary trends, researchers, research institution, and keyword distributions-to delineate the current status of Mongolian-Chinese machine translation. This study further systematically reviews research progress and emerging trends in Mongolian-Chinese MT, aiming to provide actionable insights and methodological frameworks for enhancing machine translation in analogous low-resource language pairs.

2. RESEARCH METHODS AND DATA SOURCES

The data for this study were obtained through a systematic search and screening process in the China National Knowledge Infrastructure (CNKI) database. To ensure the validity and reliability of the research data, a rigorous screening procedure was applied to select relevant literature, as follows: 1) All literature searches were completed on the same day (September 20, 2024), with the time span set to "all years"; 2) A theme-based retrieval method was adopted, using "Mongolian-Chinese machine translation" and "Chinese-Mongolian machine translation" as search terms to identify articles published in core journals (including CSSCI, CSCD, and PKU Core Journals) within the database; 3) For the retrieved literature, an in-depth manual analysis of titles, keywords, and abstracts was conducted to exclude non-research articles and studies with low relevance to Mongolian-Chinese machine translation. Subsequent steps included data cleaning, removal of duplicates, and final confirmation of 109 valid articles published between 2002 and 2024; 4) After processing the bibliographic information of the selected literature, the data were imported into CiteSpace software for visual analysis.

3. RESEARCH RESULTS AND ANALYSIS

3.1 Research Trends Analysis

Statistical analysis of publication volume and temporal distribution in the literature reflects the research dynamics and growth trends within a field, serving as a critical basis for assessing scholarly progress. After a series of data statistics and analysis, it is found that machine translation research shows significant time discrepancies and domain imbalances between China and foreign academia. The exploration of machine translation by the international academic community emerged during the Cold War era, driven by technological competition. The earliest record of research on machine translation in the Web of Science core database is the paper titled "*Interlingual Machine Translation*" published in 1958 [4], marking the beginning of systematic research in this field. Academic circles in China also initiated relevant research during the same period. According to the CNKI database, the 1959 publication *On Language Analysis Issues in Russian-Chinese Machine Translation* by the Russian Department's machine translation research group of Beijing Foreign Studies University marked the beginning of domestic machine translation research in China [5]. In the same year, Liu Yongquan provided an overview of the development status of domestic machine translation research in his article *Progress in China's Machine Translation Rese*arch [6]. In his article, Liu pointed out that "machine translation research began formal work in 1958... mainly focusing on Russian-Chinese and English-Chinese machine translation" [6], indicating the timeliness of China's introduction of translation technology. However, compared with the rapid development of research on mainstream languages, machine translation studies involving ethnic minority languages demonstrated a noticeable lag. The first academic achievement in Mongolian-Chinese machine translation did not emerge until 2002, with Badamaaodesier's "*Description of Grammatical Attributes of Mongolian Words in Chinese-Mongolian Machine Translation*" [7]. The paper introduced the grammatical attributes of Mongolian words in the Chinese-Mongolian machine translation system and their implementation in lexicographic databases. It constructed a classification system for Mongolian word classes and phrases, innovatively designing grammatical attribute fields and their annotation specifications, thereby laying critical theoretical and technical

foundations for subsequent system development. A technical path combining rule bases with dictionaries was adopted, with an initial focus on translating government work reports.

According to the data analysis in "Figure 1", the field was in its initial exploratory stage between 2002 and 2008, with an average of 0.57 publications per year. Over this seven-year span, only four research achievements were achieved, indicating a relatively small research scale and limited academic impact. During this stage, Mongolian-Chinese machine translation remained in its infancy, primarily focused on theoretical exploration and technical feasibility verification. The slow development trend can be attributed to two main factors: 1) the inherent complexities of ethnic minority language processing (morphological richness and syntactic flexibility in Mongolian); 2) insufficient allocation of academic resources to non-dominant language technologies at that time. Compared to the groundbreaking advancements in mainstream language pairs—such as statistical machine translation models for English-Chinese systems—Mongolian-Chinese research remained relatively peripheral. However, with the popularity of deep learning technology, this development gap gradually narrowed.

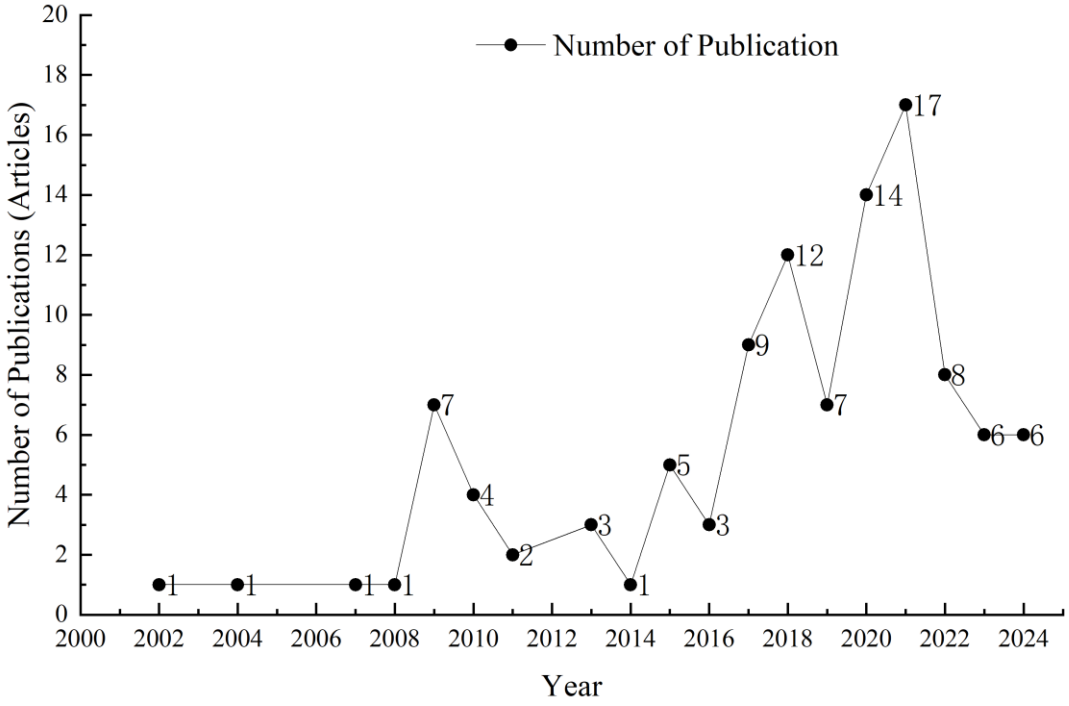

Figure 1 Annual distribution of publications on Mongolian-Chinese machine translation.

In 2009, Mongolian-Chinese machine translation research began to attract attention, with a gradual increase in related studies, reaching in an annual publication count of 7 articles. However, between 2010 and 2014, the number of studies decreased, dropping from 4 articles to 1 article, signalling unresolved challenges during this period. After 2015, driven by advancements in statistical machine translation (SMT) technologies, the number of studies began to rise again, reaching 9 articles by 2017. This trajectory shows significant fluctuations in Mongolian-Chinese machine translation studies between 2009 and 2017. Dominant research during this period focused on statistically based Mongolian-Chinese machine translation. Scholars conducted comprehensive technical explorations tailored to Mongolian linguistic characteristics, including multiple aspects such as morpheme mediation, part-of-speech tagging, and post-processing of translations, and delved into technical methods in areas like phrase-level translation, word order adjustment, and quantifier translation. Despite the progress made, domestic research still requires further strengthening in theoretical refinement and technological innovation to address persistent challenges in this specialized field.

As depicted in "Figure 1", the year 2018 witnessed a surge to 12 publications, marking the transition of Mongolian-Chinese machine translation research in China into a phase of rapid

and stable development. This progression can be primarily attributed to groundbreaking advances in neural machine translation (NMT), which provided novel methodological frameworks for low-resource language pairs. From 2018 to the present, cumulative publications have reached 70, with an annual average of 10 papers—a 105.88% increase compared to the preceding period. Notably, between 2018 and 2022, for five consecutive years, the number of publications remained above seven (with an average of 11.6 per year), peaking at 17 in 2021. Despite minor fluctuations in subsequent years (6 papers annually from 2023 to 2024), output remains relatively stable at a high level. Research during this era predominantly focused on Mongolian-Chinese NMT systems. Research during this era predominantly focused on Mongolian-Chinese NMT systems. Neural network models demonstrated remarkable performance in handling large-scale aligned corpora translation tasks, particularly excelling at translating complex syntactic structures and non-canonical linguistic phenomena. These technological advancements spurred further exploration of translation for low-resource language pairs among relevant scholars. To address critical challenges in Mongolian-Chinese machine translation such as data sparsity, limited vocabulary size, and overfitting in Mongolian-Chinese translation, researchers have employed various methods in their experiments to enhance the performance of Mongolian-Chinese machine translation systems. These methods include adversarial learning, dual learning, transfer learning, data augmentation, and the application of pre-trained models.

3.2 Researchers Analysis

By utilizing Citespace to analyze researcher collaboration patterns, one can identify the geographical distribution of core scholars and the composition and cooperative relationships of research teams in the Mongolian-Chinese machine translation (MCMT) field. Within this domain, four scholars have published over 10 papers each: Su Yi-la and Renqing Dao-er-ji from Inner Mongolia University of Technology, Hou Hongxu from the School of Computer Science at Inner Mongolia University, and Wang Siriguleng from Inner Mongolia Normal University. Among them, Su Yi-la leads with 20 articles, followed by Renqing Dao-er-ji with 17, while the other two scholars each contributed 10 articles. An additional six researchers have published more than 5 papers in this field. All aforementioned scholars focus primarily on MCMT research. Hou Hongxu initiated work on Mongolian-Chinese mixed text processing systems as early as 1994. and has been dedicated to the research and development of Mongolian-Chinese machine translation ever since, with research spanning natural language processing, artificial intelligence, and information retrieval. Suyila's research concentrates on artificial intelligence, machine learning, and web intelligence, having led significant projects including "Corpus-Based Mongolian-Chinese Machine Translation Research" and "Key Technologies for Mongolian Semantic Web." Renqing Dao-er-ji specializes in data mining, optimization, and intelligent algorithms. According to Zhang Renran's (2024) CNKI statistical analysis [3], while the machine translation field contains 1,289 core articles, MCMT-related research accounts for only 109 papers (8.4% of total). This notable disparity reveals that MCMT, despite being a crucial branch of machine translation, demonstrates substantially lower research activity compared to the broader field, indicating substantial untapped research potential and development opportunities.

This study utilized CiteSpace software to construct a core author collaboration network diagram for Mongolian-Chinese machine translation research, employing authors as the node type. The results are illustrated in Figure 2. The analysis report generated indicates 111 author nodes connected by 237 lines, with a network density of 0.0388 and a Q value (modularity) of 0.7367. As noted in reference, "a Q-value exceeding 0.3 indicates a statistically significant community structure in the network, and values approaching 1 reflect superior clustering efficacy."[8] "Figure 2" demonstrates that high-productivity authors exhibit densely interconnected nodes, underscoring their substantial influence in the field and frequent collaborative interactions. However, most authors display sparse and scattered connections, suggesting that MCMT research remains concentrated among a few scholars, while broader collaborative relationships remain fragmented. This highlights the need to foster multi-tiered, cohesive research teams and enhance the frequency and intensity of interdisciplinary collaborations to strengthen the overall research framework.

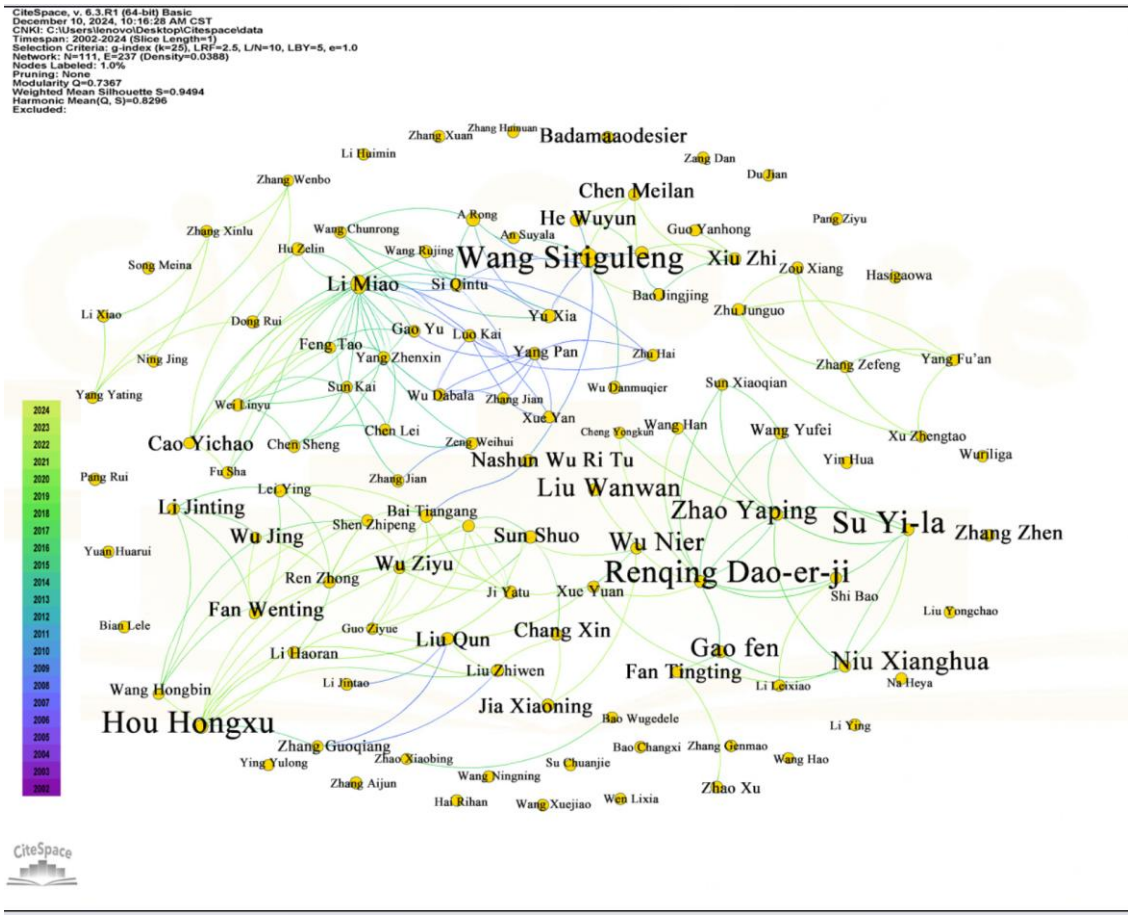

Figure 2 Researcher collaboration network diagram.

3.3 Research Institution Analysis

An analysis of research institutions in Mongolian-Chinese machine translation (MCMT) using CiteSpace generated “Figure 3” and “Table 1”. The results reveal that the 109 core publications originated from 30 institutions, predominantly higher education institutions, research institutes, and key laboratories. Among the top eight most productive institutions, eight are computer science and information engineering schools located in Inner Mongolia, with only two situated outside the region. The majority of publications first appeared in 2009, indicating that MCMT research gained broader regional attention in China during this period, fostering cross-regional collaboration. From the perspective of inter-institutional collaboration, among the 30 nodes in the current research system, there are only 21 connections, with a node density of 0.0483. This suggests that research in this low - resource domain is geographically limited, with insufficient interaction and knowledge exchange among institutions. Therefore, promoting cross-institutional collaboration to boost overall research efficiency is both urgent and necessary.

Figure 3 Network diagram of cooperative institutions.

Table 1. Number of publications among top 8 prolific organizations and the year of their first publication

Institution Name	Publication Volume	First Publication Year
Inner Mongolia University of Technology	38	2016
Inner Mongolia University	37	2009
Inner Mongolia Normal University	24	2010
University of Science and Technology of China	9	2009
Institute of Intelligent Machines，Chinese Academy	6	2009
Chinese Academy of Sciences	5	2007
Kunming University of Science and Technology	2	2022
Office of the Yunnan Provincial Committee for the Guidance of Ethnic Minority Languages and Scripts	2	2020

3.4 Research Hotspot Analysis

3.4.1 Keyword Co-occurrence Mapping Analysis

Keywords serve as concise summaries of the thematic focus of scholarly literature. Analyzing high-frequency keywords can reveal research hotspots, trends, and interconnections among research topics within a field [9]. Using CiteSpace, we conducted a keyword analysis of core publications, generating a visual network map comprising 112 keyword nodes and 143 connection lines. The network exhibits a density of 0.023 and a modularity (Q-value) of 0.7367, exceeding the 0.3 threshold, which confirms the statistical significance and rationality of the clustering results. These findings are presented in “Figure 4” (“Keyword Co-occurrence Network”) and “Table

2" ("High-Frequency Keyword Statistics"). Based on the analyzed data, the top ten keywords ranked by frequency from highest to lowest are: "machine translation", "Mongolian language", "Mongolian script", "neural networks", "pre-training", "language models", "out-of-vocabulary words", "translation models", "Mongolian-Chinese translation", and "reverse translation". These results indicate that current research on Mongolian-Chinese machine translation (MCMT) predominantly focuses on neural network-based approaches. Scholars are dedicated to continuously optimizing language and translation models to enhance translation quality and system performance. Key methodologies include pre-trained model techniques and reverse translation methods based on data augmentation. Notably, interdisciplinary connections—such as "morphology", "word reordering", and "artificial intelligence" — demonstrate the expanding scope and deepening integration of cross-domain insights in MCMT research.

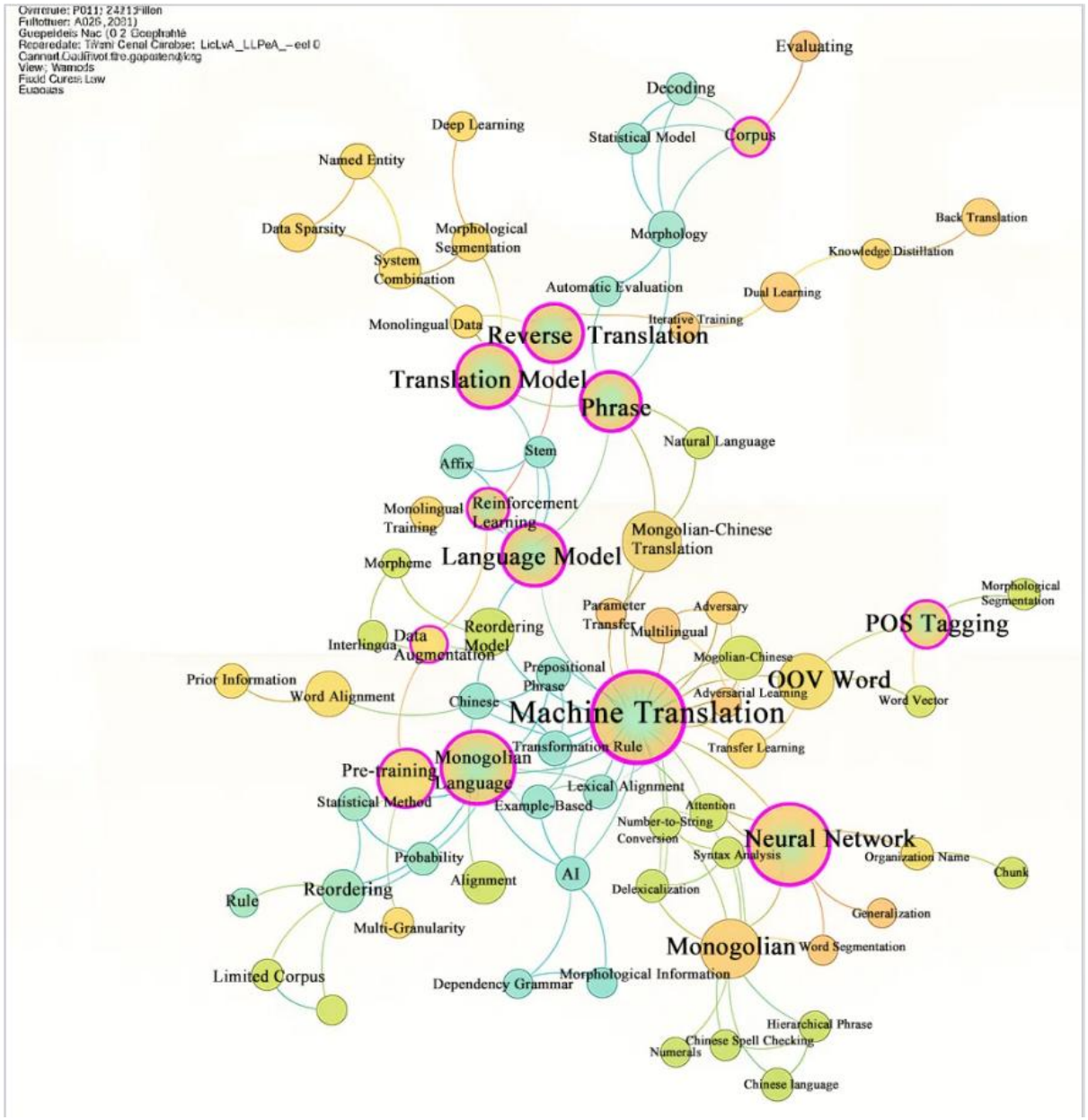

Figure 4 Keyword co-occurrence graph.

Table 2. Top 10 high-frequency keywords

Keywords	Frequency	First Publication Year
machine translation	19	2004
Mongolian	17	2004
neural network	6	2017
pre-training	5	2019
Language model	5	2008
out-of-vocabulary words	5	2017
Translation model	4	2011
Mongolian-Chinese translation	4	2019
reverse translation	4	2020
phrase	4	2010

3.4.2 Keyword Clustering Map Analysis

Keyword clustering analysis enables a deeper exploration of research hotspots and thematic focuses in Mongolian-Chinese machine translation (MCMT). This study configured the time slicing parameter as "January 2000 to December 2024" to ensure data comprehensiveness and methodological rigor. The year-per-slice was set to "1", treating each year as an independent temporal unit, with "Keyword" selected as the node type and "g-index" applied as the primary filtering criterion. Using the log-likelihood ratio (LLR) algorithm for clustering, six distinct clusters (#0–#6) were generated (see "Figure 5"), clearly showing the distribution of research topics in this field. The modularity Q-value of 0.7367 is well above the 0.3 threshold, indicating a significant cluster structure. Additionally, the silhouette score (S value) of 0.9494, well above the 0.7 benchmark, attests to the high credibility of cluster labeling. The six clusters, labeled based on keyword analysis, are: #0 Chinese language, #0 Chinese, #1 Mongolian script, #2 Dual Learning, #3 System Combination, #4 Morphology, #5 Statistical Methods, #6 Out-of-Vocabulary Words. These clusters reflect domains with high co-citation frequencies in MCMT research. By synthesizing these clustering results with high-frequency and high-centrality keywords, the core research themes in MCMT can be further distilled, providing a structured framework for understanding interdisciplinary priorities and methodological trends.

First research on the linguistic characteristics and morphological information processing of Mongolian and Chinese languages has long been a focal point in this field. This theme encompasses clusters #0, #1, #4, and #5, with core keywords including "Chinese language", "morphological information", "tree-to-string model", "Mongolian script", "morphology", "phrases", "statistical methods", and "probability". These clusters indicate that scholars in Mongolian-Chinese Machine Translation (MCMT) have consistently concentrated on the differences in linguistic features between Mongolian and Chinese. Mongolian, as an agglutinative language, exhibits rich morphological variations, whereas Chinese, an isolating language, lacks morphological changes. There are significant differences between the two in grammar, lexicon, and other aspects. Therefore, it is essential to study the correspondence of Mongolian-Chinese vocabulary, as well as the analysis of Mongolian syntactic structures. The construction of translation models requires consideration of issues related to morphological segmentation and restoration. To mitigate data sparsity caused by morphological differences and the scarcity of large-scale parallel corpora, Yang Pan et al. (2009) introduced morphological analysis into statistical machine translation factorized models, effectively alleviating challenges such as morphological mismatches and disordered word selection in Chinese-Mongolian translation [10]. Luo Kai et al. (2009) incorporated source-language dependency syntax information and target-language morphological information in Chinese-Mongolian translation model, leveraging lexicalized dependency syntactic features and extracting Mongolian morphological information to adapt factored models using the LOP principle. Experimental results demonstrated that their method achieved significantly higher BLEU scores compared to traditional phrase-based SMT models [11].

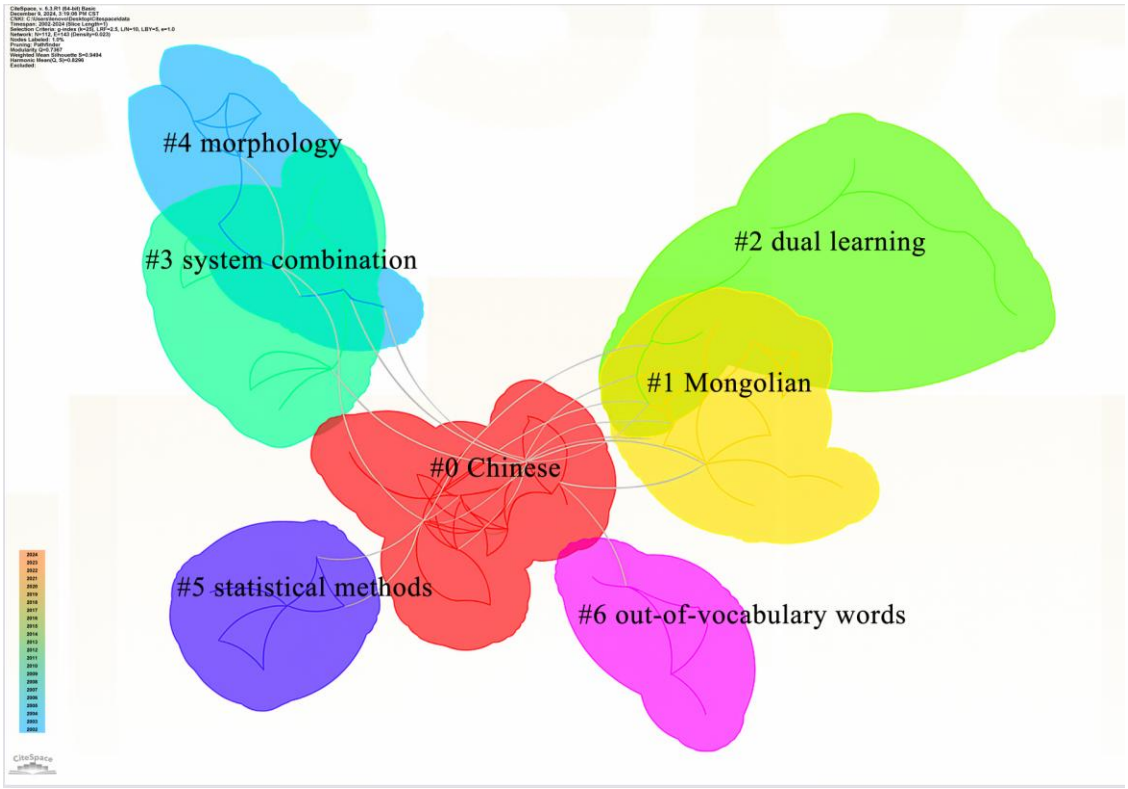

Figure 5 Clustering knowledge atlas of keywords.

Neural network-based translation models have emerged as the dominant paradigm in Mongolian-Chinese machine translation (MCMT). This theme is covered in clusters #1, #3, and #6, with core keywords including “neural networks”, “BiLSTM-CNN-CRF”, “syntactic analysis”, “named entities”, “unlisted word”, “part-of-speech tagging”, and “system integration”. Additionally, the translation model, particularly the Transformer architecture, lies at the heart of machine translation research. MCMT has evolved from early Rule-Based Machine Translation (RBMT), Example-Based Machine Translation (EBMT) to Statistical Machine Translation (SMT), and now to Transformer-based large language models (LLMs). Each iteration has refined and optimized prior frameworks, driving marked improvements in model performance and facilitating steady advancements in low-resource machine translation quality. Key challenges in MCMT include data sparsity, the agglutinative nature of Mongolian, limited vocabulary size, low word frequency, and OOV word handling. To address these issues, the academic community has progressively developed innovative methodologies and achieved systematic research progress, transitioning from early statistical-based models to modern deep learning-driven architectures.

Statistical analyses of literature data reveal that recent advancements in Mongolian-Chinese machine translation have predominantly focused on the following methodologies: adversarial learning mechanisms, dual learning (integrating semi-supervised learning), cross-lingual transfer learning strategies, reverse translation methods based on data augmentation, and applications of pre-trained language models. These approaches are reflected across clusters #1, #2, and #6. Scholar He Wuyun (2022) conducted focused research on the application of diverse deep learning models to Mongolian word segmentation. The study comprehensively analyzed the impact of different segmentation methods on Transformer-based Mongolian-Chinese machine translation model and proposed an improved neural network Mongolian word segmentation method based on this analysis. Experimental results showed that the refined BiLSTM-CNN-CRF segmentation model achieved exceptional performance, with an accuracy rate of 97.37% and an optimal BLEU-5 score of 73.30% [12]. This advancement provides a critical reference for improving translation quality in low-resource MCMT systems.

3.4.3 Keyword Burst and Timeline Diagrams Analysis

Keyword burst analysis enables a clear visualization of the historical evolution of research domains. By chronologically organizing the emergence timelines of keywords and presenting the attention cycles of research hotspots from past to present, it reveals the developmental trajectory and frontiers of the field. To systematically visualize this process, this study selected the top 15 keywords with the highest burst intensity in Mongolian-Chinese machine translation research, conducted quantitative statistical analysis, and subsequently generated a timeline diagram ("Figure 6") and a keyword burst diagram ("Figure 7"). These diagrams intuitively reflect the evolutionary trends of research hotspots and the dynamics of scholarly attention. Analysis of these two diagrams is summarized as follows: Mongolian-Chinese machine translation emerged in China starting from 2002, coinciding with the preliminary exploration phase of machine translation technology. Initial research primarily focused on developing rule-based and example-based Chinese-Mongolian machine translation systems. Against the backdrop of machine translation advancements, domestic research on Mongolian-Chinese machine translation gradually commenced between 2002 and 2008. In 2007, scholars Hou Hongxu et al. developed an experimental Chinese-Mongolian EBMT system based on word alignment through experimental research, marking a paradigm shift from rule-based approaches to corpus-based methodologies in Mongolian machine translation[13]. Despite methodological breakthroughs, challenges persisted due to the linguistic characteristics of Mongolian, errors occurred in word form processing, and the evaluation method for generated results was overly simplistic. Between 2009 and 2017, Mongolian-Chinese machine translation research exhibited fluctuating development, primarily centred on statistical machine translation (SMT) frameworks. The research primarily focused on three model types: word-based, phrase-based, and syntax-based approaches. Scholars, including Wang Siriguleng (2011), proposed a Chinese source sentence reordering method based on Mongolian target word order. By constructing a syntactic reordering rule repository and designing a binary tree reordering algorithm, this approach preemptively adapts Chinese word order to align with Mongolian linguistic conventions. Experimental results on the CWMT2009 corpus demonstrated that the method improved the system's BLEU score by 2.07% (development set) and 0.37% (test set) compared to baseline models, significantly enhancing the performance of Chinese-Mongolian machine translation systems at the time. This advancement provided a novel research paradigm for optimizing grammar-based statistical translation models [14]. The keywords shown in the figure—Mongolian, language model, reordering, phrases, and translation model—highlight the research priorities during the period of statistical-based Mongolian-Chinese machine translation.

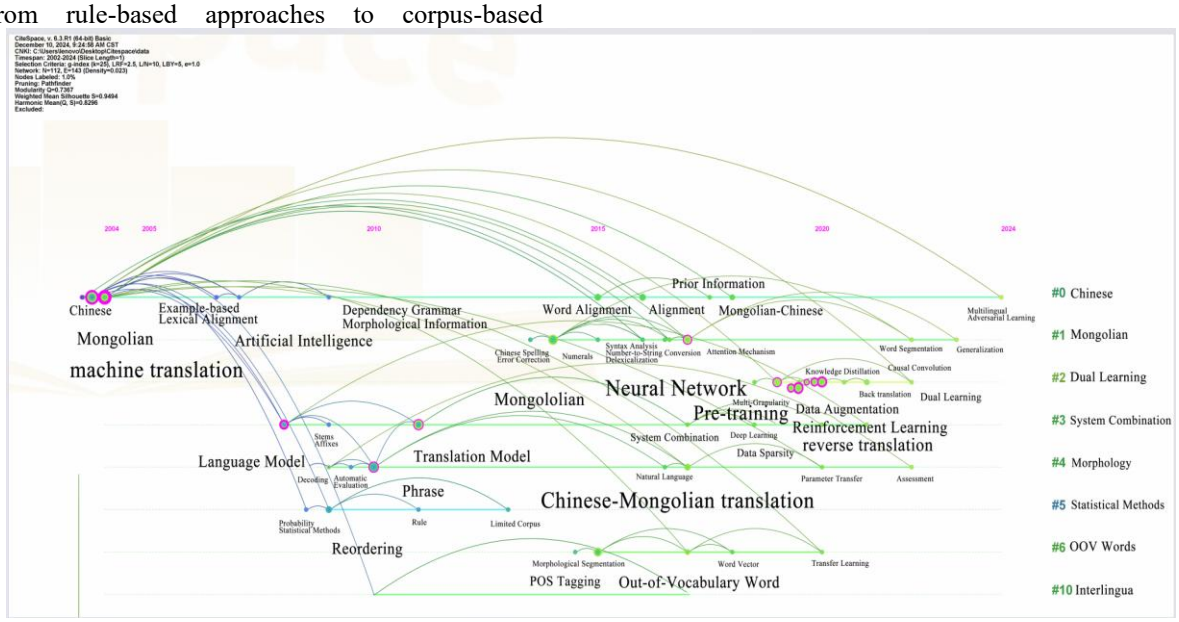

Figure 6 Keyword timeline chart.

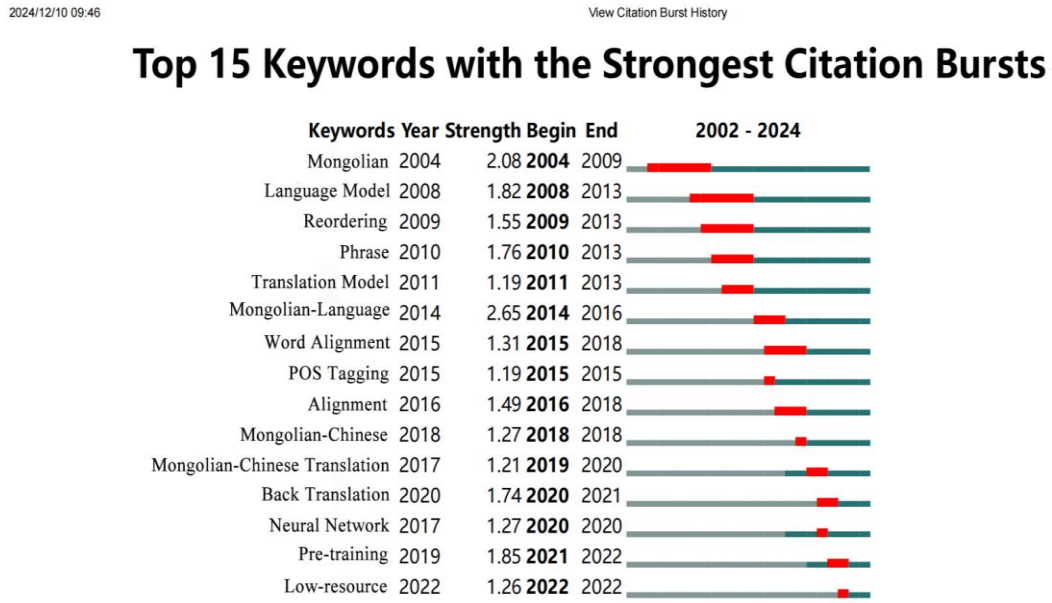

Figure 7 Top 15 keywords with the strongest citation bursts.

Since 2018, Mongolian-Chinese machine translation, primarily based on Neural Machine Translation (NMT), has entered a phase of rapid and stable development. Proposed in 2014, NMT leverages neural network architectures from deep learning to automate translation from source to target languages and is widely used in various machine translation. Major NMT models include Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Transformer models. In the Mongolian-Chinese machine translation field, Hou Hongxu et al. (2022) highlighted in their review that the main factors affecting system performance stem from language pair scarcity and Mongolian's agglutinative nature[15]. The adoption of neural network models has created new possibilities for improving translation quality. Liu Wanwan et al. (2018) proposed a Mongolian-Chinese translation model integrating part-of-speech (POS) tagging with a gated recurrent unit (GRU) network and global attention mechanisms. Experimental results demonstrated that this approach significantly outperformed traditional statistical methods and baseline RNN systems, particularly in enhancing long-sentence processing capabilities [16]. In 2023, Xue Yuan et al. [17] developed a graph convolutional network (GCN)-based approach that directly encodes dependency syntax trees using directed graph encoding to reflect node relationships, while feeding high-quality predicted translations back into training. Experiments on a Mongolian-Chinese parallel corpus of 1.2 million sentence pairs revealed BLEU score improvements of 2.69 and 2.09 over conventional bidirectional RNN (BiRNN) and CNN models, respectively, validating the method's efficacy in enhancing translation performance. In a deeper analysis of neural machine translation, Zhang Junjin et al. (2024) focused on the application and research of neural machine translation in low-resource languages, summarizing that "Current mainstream Transformer-based NMT frameworks have driven significant progress in neural machine translation for low-resource or zero-resource languages, model architectures, and automatic summarization"[18]. Collectively, these studies underscore the tangible advancements achieved in Mongolian-Chinese neural machine translation research.

Further analysis through keyword emergence trends reveals that in the field of Mongolian-Chinese machine translation research over the past five years (2020-2024), "reverse translation", "neural networks", "pre-training", and "low-resource" have emerged as focal point of scholarly interest, with sustained growth in related research attention. Hou Hongxu et al. (2022) indicate that in contemporary Mongolian-Chinese neural machine translation studies, reverse translation technology for generating pseudo-parallel corpora has become the primary means of corpus expansion. This approach effectively mitigates data sparsity issues while enhancing model robustness. Bai (2020) proposed a data augmentation method using reverse

translation technology to improve the performance of Mongolian-Chinese machine translation models. This method constructs a Chinese-Mongolian translation system to generate Mongolian-predicted sentences from Chinese monolingual corpora, which are then concatenated with the original Chinese sentences to form synthetic bilingual corpora. Finally, the synthetic corpora are combined with real parallel corpora at a certain ratio for training the Mongolian-Chinese machine translation model. Experimental results show that when the ratio of pseudo corpora to real corpora is 1:1, the model performance improves most significantly [19].

In the realm of pre-trained model research, Su Yila et al. (2021) proposed a method for pre-training cross-lingual Mongolian-Chinese language models based on self-attention mechanisms, using monolingual corpora combined with a small amount of parallel corpora to improve Mongolian-Chinese machine translation performance. Experimental results confirmed this approach significantly enhances translation quality, validating the effectiveness of pre-trained language models in Mongolian-Chinese machine translation[20]. In addition to methods aimed at improving Mongolian-Chinese machine translation quality, low-resource machine translation has gradually attracted attention from scholars in the field. The evolution of research on Mongolian-Chinese machine translation not only demonstrates practical methods for enhancing translation quality for specific language pairs but also provides valuable research ideas and optimization solutions for other language combinations facing similar challenges, thus driving the development of the low-resource machine translation filed.

4. CONCLUSION AND PROSPECT

4.1 Conclusion

Analyzing from the quantity of published articles, Mongolian-Chinese machine translation research over the past two decades has exhibited distinct phase characteristics: Initial Exploration Phase (2002-2008), the field was in the exploratory stage, with limited research output and minimal influence. The annual average publication count stood at 0.57 articles(total of 4 articles). Fluctuating Growth Period (2009-2017), annual productivity increased to 4.25 articles (34 total). A notable peak emerged in 2009 with 7 articles, followed by a slowdown (2010-2014): annual average of 2.5 articles). After 2015, as statistical machine translation (SMT) technologies matured, publication output resumed growth (2015-2017: annual average of 5.67 articles). Stable Development Phase (2018-2024), driven by neural machine translation (NMT) advancements, the field entered a high-quality development phase. Annual publications surged to 10 articles (total of 70 articles), representing a 105.88% increase compared to the previous period. Between 2018 and 2022, output remained consistently above 7 articles annually (annual average: 11.6), peaking at 17 articles in 2021. Although there were minor fluctuations afterward, productivity stabilized at a high level (6 articles annually in 2023-2024). During this phase, research paradigms matured, forming a sustainable academic output mechanism with stable, high-volume productivity.

Analyzing from the perspective of the researchers group, the core intellectual force in Mongolian-Chinese machine translation is primarily composed of scholars from institutions such as Inner Mongolia University of Technology, Inner Mongolia University, and Inner Mongolia Normal University. Key contributors include researchers like Su Yila, Ren Qingdaorji, Hou Hongxu, and Wang Siriguleng—all affiliated with universities in Inner Mongolia—who have exerted significant academic influence in this field, having published 20, 17, and 10 academic papers, respectively,with deep research in key technologies such as natural language processing, artificial intelligence, and information retrieval. Notably, Su Yila, with expertise in artificial intelligence, machine learning, and network intelligence, has led major projects, including "Research and Implementation of Mongolian-Chinese Statistical Machine Translation Based on Deep Learning," significantly advancing the technological development and practical application of Mongolian-Chinese machine translation. Hou Hongxu, as a senior researcher in this field, has long been dedicated to the research and development of Mongolian-Chinese machine translation systems. However, the research community in Mongolian-Chinese machine translation is relatively small compared to the broader field of machine translation and needs further expansion to enhance collaborative innovation and interdisciplinary expertise integration.

Analyzing research institutions, these institutions exhibit diversity and regional concentration, involving 30 units such as

universities, research institutes, and key laboratories. High-productivity institutions are mainly concentrated in the field of Computer and Information Engineering in the Inner Mongolia region. This is partly due to the urgent demand for Mongolian-Chinese machine translation technology in this area, and partly related to national policy support. Despite the considerable number of participating institutions, disparities exist in their initiation timelines, with most high-productivity institutions commencing their research around 2009. However, inter-institutional collaboration and communication remain relatively limited. Future research should strengthen inter-institutional collaboration and interdisciplinary cooperation while fostering collaborations with experts and scholars in linguistics, translation studies, and related fields. Such efforts will broaden research horizons by integrating technical advancements with linguistic theories, thereby advancing Mongolian-Chinese machine translation research and enhancing the accuracy, fluency, and practical applicability of translation outcomes.

Analyzing from research hotspots and frontiers, neural network-based methods currently dominate Mongolian-Chinese machine translation, giving rise to diverse neural machine translation (NMT) models. Scholars continue to explore and refine methodologies, with a primary focus on neural network—particularly the Transformer model—which has been widely adopted due to its powerful representational capabilities and computational efficiency. Such models excel at capturing deep semantic relationships between languages, significantly enhancing translation quality. Investigations into adversarial learning, transfer learning, dual learning, reverse translation techniques, and pre-trained language models have provided effective pathways for improving translation performance. Nevertheless, a critical challenge remains: how to further optimize pre-trained models to better adapt to low-resource languages like Mongolian. Strategies for enhancing parallel corpora through data augmentation methods, particularly reverse translation, constitute another focal area of current research. Additionally, refining pre-trained models to address morphological complexity, syntactic diversity, and domain-specific lexical gaps in Mongolian continues to demand targeted research and innovation.

4.2 Prospect

4.2.1 Developing OpenQA Systems

At present, the development of open question answering systems (OpenQA) in low-resource conditions represents a significant and promising research direction. Existing QA systems are primarily limited to English, while non-English languages face development constraints due to the lack of large-scale annotated datasets. Therefore, developing efficient and cost-effective open-domain QA systems holds practical significance in resource-scarce contexts. A notable example is the successful implementation of a Turkish-language OpenQA system, which offers valuable insights for other low-resource languages. This system was developed by adapting and retraining the existing ColBERT-QA system using weak supervised learning and a machine translation-generated annotated dataset SQuAD-TR, combined with unstructured knowledge sources related to the target language, namely Wikipedia. Additionally, the study evaluated the system's performance using a small number of labeled samples, validating the effectiveness of this approach in improving the performance of OpenQA systems (Emrah Budur 2024)[21]. Similarly, for the Mongolian-Chinese language pair, researchers could build target - language QA datasets through machine translation, refine model architectures using localized knowledge sources, optimize training strategies, and implement robust evaluation protocols to develop effective OpenQA systems.

4.2.2 Multimodal Mongolian-Chinese Machine Translation

Multimodal Mongolian-Chinese machine translation remains an underexplored field. As artificial intelligence technologies advance, integrating multimodal resources can address dataset limitations in low-resource environments by combining speech, video, text, and cultural symbols to enrich Mongolian-language datasets and broaden the application scope of translation models. The first multimodal dataset constructed for the English-Indic language pair was realized through video-guided multimodal machine translation (VMMT) models in low-resource conditions. The VMlM'T system employs spatio-temporal video context as an additional input modality along with the source text. The spatio-temporal video context is extracted using a pre-trained 3D convolutional neural network. Experimental results show that, for

the English-to-Indic, BLEU scores improved by +4.2 points and chrF scores by +0.07; for the Indic-to-English, BLEU scores increased by +5.4 points and chrF scores by +0.07 (Loitongbam 2023)[22]. In addition, integrating speech recognition technologies could optimize spoken language translation tasks and improve user experiences in practical applications. In low-resource speech recognition research, end-to-end automatic speech recognition (E2E-ASR) systems leveraging deep learning exhibit considerable potential. A non-autoregressive neural network-driven text-to-speech (TTS) engine efficiently converts phoneme sequences into high-quality speech waveforms (Mel spectrograms). Furthermore, Real-time data augmentation techniques applied to these spectrograms enable feature extraction for training convolutional neural network (CNN) and bidirectional long short-term memory (BLSTM)-based E2E-ASR systems. Experimental results indicate marked improvements in low-resource speech recognition, with word error rate (WER) reduced by 20.75% and character error rate (CER) by 10.34% (Sami Dhahbi 2025)[23]. Applying this technology to Mongolian-Chinese machine translation could significantly enhance the performance of spoken language translation tasks between these language pairs.

4.2.3 Model Optimization

Efficient optimization methods should be explored based on the characteristics and advantages of different models, with a focus on fusing multi-model outputs to enhance the performance of Mongolian-Chinese machine translation. The application potential of pre-trained language models (PLMs) in multilingual neural machine translation (MCNMT) can be fully explored. For instance, the time cost of model training can also be reduced. The ReMixup-NMT method, which is based on regularized Mixup for BERT knowledge fusion, by constraining distribution consistency between standard Transformer encoders and Mixup-based Transformer encoders, efficiently distills pre-trained BERT knowledge into Seq2Seq architectures without requiring extra parameter training. Evaluations on the IWSLT'15 English-Vietnamese and IWSLT'17 English-French datasets demonstrate its superiority over existing BERT fusion and dropout-based methods (Zhang, 2025)[24]. Recently, the MoE-LLM framework has emerged as an innovative approach by integrating sparse Mixture-of-Experts (MoE) components into large language models (LLMs) to enhance multilingual translation capabilities. This method employs a hybrid transfer learning strategy, freezing LLM parameters to prevent catastrophic forgetting while incorporating specialized translation expert modules. By using LLM representations for pre-warming initialization of MoE parameters, it bridges the representation gap between different languages. Experimental results reveal that MoE-LLM outperforms directly fine-tuned LLMs across 10 translation directions, achieving up to 2.5 BLEU point improvements. It also excels in zero-shot translation scenarios, surpassing strong baselines like Adapter and LoRA (Zhu, 2025)[25]. The multilingual translation capability of MoE-LLM and its optimization for low-resource scenarios effectively alleviate the issue of scarce parallel corpora. Therefore, these advantages highlight its significant potential for Mongolian-Chinese machine translation applications.

REFERENCES

[1] Li Xiang, Gao Zhaoyang. Knowledge Graph and Development Trends in Foreign Machine Translation Research [J]. Shanghai Journal of Translators, 2024, (02): 41-47.

[2] Fu Linling, Liu Lei. Visualized Analysis of Researches on Machine Translation Based on CiteSpace [J]. HEILONGJIANG SCIENCE, 2023, 14 (15): 1-5, 35.

[3] Zhang Renran, Liu Jianzhu. Visual Analysis of Intelligent Translation Research Based on CiteSpace: Retrospect and Prospect [J]. Overseas English, 2024, (07): 46-49.

[4] Richens, R. H. Interlingual machine translation [J] The Computer Journal, 1958, 1(3): 144-147.

[5] Russian Department's Machine Translation Research Group of Beijing Foreign Studies University. On Language Analysis Issues in Russian-Chinese Machine Translation [J]. FOREIGN LANGUAGE TEACHING AND RESEARCH, 1959, (06): 365-374.

[6] Liu Yongquan. Progress in China's Machine Translation Research [J]. Chinese Science Bulletin, 1959, (17): 563-564.

[7] Badamaaodesier. Description of Grammatical Attributes of Mongolian Words in Chinese-

Mongolian Machine Translation [J]. Minority Languages of China, 2002, (04): 61-63.

[8] Li Jie Chen Chaomei. CiteSpace: Text Mining and Visualization in Science and Technology [M]. Beijing: Capital University of Economics and Business Publishing House, 2022: 105, 137.

[9] Yan Weina. Progress, hotspots, and trends in research on popular science journals in China: Visual analysis based on CiteSpace [J]. Chinese Journal of Scientific and Technical Periodicals, 2024, 35(02): 163-170.

[10] Yang Pan, Li Miao et al.. Morphology-Processing in Chinese-Mongolian Statistical Machine Translation [J]. Journal of Chinese Information Processing, 2009, 23(01): 50-57.

[11] Luo Kai, Li Miao, et al. Dependency Informed Chinese-to-Mongolian Translation Model with Morphological Information [J]. Journal of Chinese Information Processing, 2009, 23(06): 98-104.

[12] He Wuyun. Mongolian Word Segmentation Method Based on Neural Network and Its Application [D]. Inner Mongolia Normal University, 2022. Inner Mongolia Normal University

[13] Hou Hongxu, Liu Qun, Nashun Wu Ri Tu. Design of professional term knowledge map automatic recognition based on database [J]. Journal of Chinese Information Processing, 2007, (04): 65-72.

[14] Wang Siriguleng, Si Qintu, Nashun Wu Ri Tu. A Reordering Method of Chinese-Mongolian Statistical Machine Translation [J]. Journal of Chinese Information Processing, 2011, 25(04): 88-92.

[15] Hou Hongxu, Sun Shuo, WU Nier. Survey of Mongolian-Chinese Neural Machine Translation [J]. Computer Science, 2022, 49(01): 31-40.

[16] Liu Wanwan, Su Yila, WU Nier, et al.. Mongolian-Chinese Machine Translation Research Based onPart of Speech Tagging with Gated Unit Neural Network [J]. Journal of Chinese Information Processing,, 2018, 32(08): 68-74.

[17] Xue Yuan, Su Yila, Renqing Dao-er-ji , et al. Mongolian and Chinese Neural Machine Translation Based on Graph Convolutional Encoder [J]. Computer Applications and Software, 2023, 40(10): 70-75, 89.

[18] Zhang Junjin, TianYonghong, Song Zheyu,et al. Survey of Neural Machine Translation[J]. Computer Applications and Software, 2024, 60(04): 57-74.

[19] BAI T G. Mongolian-Chinese Neural Network Machine Translation Based on Reinforcement Learning [D] . Hohhot: Inner Mongolia University, 2020.

[20] Su Yila, Gao Fen, Niu Xianghua, et al. Pre-Training Cross Mongolian-Chinese Language Model Based on Self-Attention Mechanism [J]. Computer Applications and Software, 2021, 38(02): 165-170.

[21] Emrah Budur, Rıza Özçelik, Dilara Soylu, et al. Building efficient and effective OpenQA systems for low-resource languages [J]. Knowledge-Based Systems, Volume 302, 2024, 112243.

[22] Loitongbam Sanayai Meetei, Alok Singh, Thoudam Doren Singh, et al. Do cues in a video help in handling rare words in a machine translation system under a low-resource setting? [J]. Natural Language Processing Journal, Volume 3, 2023, 100016.

[23] Sami Dhahbi, Nasir Saleem, Sami Bourouis, et al. End-to-end neural automatic speech recognition system for low resource languages [J] Egyptian Informatics Journal, Volume 29, 2025, 100615.

[24] Guanghua Zhang, Hua Liu, Junjun Guo, et al. Distilling BERT knowledge into Seq2Seg with regularized Mixup for low-resource neural machine translation [J]. Expert Systems with Applications. Volume 259, 2025, 125314.

[25] Shaolin Zhu, Leiyu Pan, Dong Jian, et al. Overcoming language barriers via machine translation with sparse Mixture-of-Experts fusion of large language models [J]. Information Processing & Management, Volume 62, Issue 3, 2025, 104078.