Research on Core Issues and Mainstream Algorithms of Chinese Word Segmentation

Yuemeng Ren

School of Intelligent Science and Engineering, Chengdu Neusoft University, Sichuan, China
Email: 594251028@qq.com
Manuscript received May 1, 2026; accepted May 13, 2026; published June 15, 2026

***Abstract*—With the rapid development of large language models and the deep integration of artificial intelligence into various industries the requirements for accuracy and generalization ability in Natural Language Processing (NLP) are constantly rising. Large models centered on Transformers are propelling NLP into a new stage. Chinese lacks natural word boundaries, making word segmentation a fundamental step in Chinese NLP, and its accuracy directly determines the effectiveness of subsequent tasks. Due to its linguistic characteristics, Chinese word segmentation has long faced three core challenges: inconsistencies between general vocabularies and segmentation standards, difficulties in handling ambiguous segments, and poor performance in out-of-vocabulary word identification. The paper highlights the advantages of deep learning for word segmentation, detailing classic neural networks such as CNN, RNN, LSTM, and BiLSTM-CRF, as well as the application of pre-trained models including BERT, RoBERTa, and lightweight real-time models. The paper emphasizes the advantages of deep learning in word segmentation, detailing classic neural network models such as CNN, RNN, LSTM, and BiLSTM CRF, as well as the application of BERT, RoBERTa pre-trained models, and lightweight real-time models in word segmentation. Research shows that deep learning-based word segmentation methods offer the best overall performance, effectively solving challenges in ambiguous segmentation and out-of-vocabulary word recognition. Different algorithms and systems can meet the diverse needs of scientific research, industry, and vertical fields. This study clarifies the evolution of Chinese word segmentation technology, providing a reference for the selection, engineering implementation, and optimization of word segmentation algorithms in the large model era, and is highly valuable for advancing the high-quality development of Chinese natural language processing.**

***Keywords*—large language model, chinese word segmentation, deep learning, natural language processing**

I. Introduction

Against the backdrop of rapid breakthroughs in large language model technology, natural language processing has become a core research direction in computer science and artificial intelligence [1]. Chinese differs fundamentally from Western alphabetic writing systems, with no explicit separators between words in sentences. Therefore, Chinese word segmentation is a prerequisite for achieving high-level language processing tasks such as semantic understanding, text mining, and machine translation. The accuracy of word segmentation directly affects the performance of various downstream high-level tasks [2].

The unique linguistic characteristics of Chinese and the complexity of its practical application scenarios have long led to several key technical challenges in Chinese word segmentation. Chinese word segmentation lacks unified segmentation standards and vocabulary specifications. The granularity of word segmentation varies significantly across different corpora and application scenarios. The compatibility between general vocabulary and domain-specific vocabulary is poor, directly leading to difficulties in interoperability between the outputs of different word segmentation systems. Furthermore, Chinese contains a large number of intersectional, combinatorial, and contextual ambiguous fields. The same string can form multiple reasonable segmentation methods in different contexts. Traditional methods rely solely on string matching or shallow statistical features, making effective disambiguation difficult. In addition, out-of-vocabulary words include various types such as personal names, place names, new internet words, professional terms and abbreviations. These words are not covered by predefined dictionaries and training corpora, and traditional word segmentation methods exhibit extremely limited recognition capability for such words. The above problems have long restricted the development and large-scale application of Chinese word segmentation technology [3].

After decades of research and iteration, Chinese word segmentation has gradually evolved from the traditional dictionary matching and statistical modeling stage to the artificial intelligence-driven deep learning stage. However, core problems such as ambiguity resolution and out-of-vocabulary word recognition have not been completely solved [4]. This paper takes Chinese word segmentation technology as the research topic, systematically sorts out relevant research results at home and abroad using the literature review method, comprehensively summarizes the traditional word segmentation methods based on dictionaries, statistics and rules, and focuses on deeply analyzing the application advantages of deep learning in word segmentation tasks, including classic neural network models such as CNN, RNN, BiLSTM CRF, as well as pre-trained models such as BERT, RoBERTa and lightweight real-time word segmentation models [5]. This paper aims to clarify the advantages, disadvantages, characteristics, and applicable scenarios of various word segmentation methods, and to outline the overall development of Chinese word segmentation technology. It provides theoretical support for the optimization, improvement, and engineering implementation of Chinese word segmentation technology in the era of large models, and has important practical significance for improving the overall application level of Chinese natural language processing.

II. Traditional Chinese Word Segmentation Methods

Traditional Chinese word segmentation methods were the

core support for Chinese information processing before the rise of deep learning technology. Their technical framework is built around three core paths: "dictionary matching" "statistical modeling" and "rule-based reasoning" laying the theoretical and engineering foundation for Chinese word segmentation. This chapter serves as a comparative reference for subsequent deep learning word segmentation methods. Adhering to the principle of simplicity, it unfolds the core logic of three traditional methods in sequence: First, it introduces dictionary-based mechanical word segmentation, outlining mainstream matching algorithms and their technical limitations; second, it elaborates on statistical word segmentation methods, focusing on the principles and characteristics of core models; finally, it briefly describes the approach and current applications of rule-based word segmentation.Each of these three methods has its own focus, collectively forming the technical landscape of traditional Chinese word segmentation. This chapter will highlight the core points and streamline redundant details, providing a comparative framework for the deep learning methods discussed in detail later.

A. Dictionary-Based Word Segmentation Methods

Dictionary-based word segmentation (also known as mechanical word segmentation) is the most basic technical solution for Chinese word segmentation. Its core idea is to match the Chinese character string to be processed with entries in a machine dictionary according to a predefined strategy; a successful match indicates completion of word segmentation. This method can be classified according to scanning direction (forward/reverse) and matching length priority (maximum/minimum). Due to the characteristic of Chinese characters forming words, minimum matching is rarely used; forward maximum matching (FMM), reverse maximum matching (RMM), and bidirectional maximum matching (BM) are the mainstream solutions.

FMM follows a left-to-right scanning logic, matching the string with the longest dictionary entry length; if it fails, the last Chinese character is deleted and the attempt is repeated. RMM scans in reverse, deleting the first Chinese character when it fails, resulting in better ambiguity resolution (error rate 1/245). BM combines the results of both and selects the best, achieving correct segmentation in 90% of sentences and significantly reducing ambiguity [6]. Furthermore, derivative algorithms such as word-by-word matching and minimum segmentation are limited in practical application due to their slow speed and poor adaptability to out-of-vocabulary (OOV) words. These methods are simple to implement and easy to engineer, but they lack semantic understanding capabilities. The effectiveness of ambiguity handling and out-of-vocabulary word recognition depends on the quality of the dictionary. Currently, they are mostly used as preprocessing modules in combination with other word segmentation methods.

B. Statistical Word Segmentation Methods

Statistical word segmentation methods eliminate the reliance on pre-defined dictionaries. The core logic is to learn the word-formation patterns of Chinese character combinations through large-scale labeled corpora (the higher the co-occurrence frequency of adjacent characters, the higher the probability that they form a word), thereby achieving the segmentation of unknown text. The basic process involves full text segmentation, calculating the probability of each segmentation path, and selecting the optimal result. The core reliance is on the construction and parameter estimation of statistical models.

Mainstream models include: N-gram models (assuming the current word depends only on the previous N-1 words; the ternary model is the most widely used, simple and easy to implement but dependent on corpus quality) [7]; Hidden Markov Models (HMM, based on a double stochastic process, inferring word boundary marker sequences from Chinese character sequences, achieving good fitting results for Chinese sequences) [8]; Maximum Entropy Models (ME, capable of fusing multiple features, strong generalization ability but complex training and prone to overfitting) [9]; and Conditional Random Fields (CRF, global feature modeling, overcoming the limitations of HMM and ME, and a representative model of statistical word segmentation) [10]. These methods are superior to mechanical word segmentation in terms of OOV word recognition and ambiguity resolution, but they are highly dependent on the quality of labeled corpora, exhibit high computational complexity, lack deep semantic understanding, and have weak interpretability in the decision-making process.

C. Rule-Based Word Segmentation Methods

The core of rule-based word segmentation methods is to replicate the logic of human language understanding, transforming syntactic and semantic knowledge into executable rules to achieve accurate processing of ambiguous fields. Its system architecture consists of a word segmentation subsystem, a syntactic-semantic subsystem, and a central control system. The central control system coordinates the word segmentation subsystem to acquire syntactic and semantic information and assists in ambiguity judgment; essentially, it is an expert system-based word segmentation method in the field of artificial intelligence [11].

This method constructs a reasoning network through explicit rules, making it knowledge-rich, easy to maintain, highly interpretable, and effective in segmenting complex ambiguous segments. However, it has significant limitations: it cannot autonomously learn new rules, and maintenance costs surge as the knowledge base expands; the generality and complexity of Chinese language knowledge make it difficult for rules to fully cover all aspects, and the processing efficiency of texts with overlapping ambiguities is relatively low. Although methods combining rules and statistics have emerged since then, such methods are still in the experimental and exploratory stage and have not yet been implemented on a large scale [12].

III. Deep Learning-Based Word Segmentation Methods

Compared to dictionary-based mechanical matching methods and statistical machine learning methods, deep learning has significant advantages in Chinese word segmentation tasks: First, it can automatically learn multi-level features such as characters, words, and context from the original text, eliminating the need for manual design of complex linguistic rules or statistical features, thereby

significantly reducing the cost of feature engineering. Second, models represented by recurrent neural networks and Transformers can achieve long-distance context modeling, more accurately capturing semantic dependencies, and their ambiguity resolution capabilities far surpass traditional methods. Third, supported by large-scale unlabeled corpora and pre-training mechanisms, it has stronger generalization capabilities and significantly enhances the recognition performance of out-of-vocabulary (OOV) words. Fourth, it can achieve joint modeling of word segmentation with tasks such as part-of-speech tagging and named entity recognition, constructing an end-to-end text processing system, achieving a qualitative leap in accuracy, robustness, and engineering feasibility. The following sections will discuss the technological evolution of deep learning word segmentation, typical neural network word segmentation models, key optimization strategies, and practical applications, systematically outlining the core technological development and trends in this field.

A. Neural Network-Based Word Segmentation Models

Early neural network word segmentation methods in the field of artificial intelligence simulated the human brain's neural system and were the prototype of deep learning word segmentation. They possessed the abilities of association, adaptation, and self-learning, and could learn Chinese character combinations and semantic rules autonomously through training, without the need for manually constructing a rule base. However, they suffered from problems such as complex model structures, long training cycles, poor interpretability, and poor performance in OOV word segmentation.

In 2006, Hinton proposed the concept of deep learning, which automatically learned deep features of data through multi-layer networks, breaking through the performance bottleneck of traditional shallow learning [13]. Deep learning originated from artificial neural networks in the 1940s, which aims to simulate the visual mechanism of the human brain; its Deep Belief Network (DBN) solved the model optimization problem, while the Convolutional Neural Network (CNN) proposed by LeCun *et al.* in 1998 was the first truly multi-layer structural learning algorithm [14].

As a typical deep learning technology, CNN has achieved significant results in speech recognition, image recognition, and other fields. Its powerful local feature extraction capabilities have attracted domestic and foreign scholars to extend it to the field of text processing, eliminating the tedious manual feature engineering by automatically extracting text features [15]. This method not only outperforms traditional algorithms in classification but also has high practical applicability, thus becoming an important technical framework in the field of text processing. Driven by deep learning technology, neural network-based word segmentation models have been iteratively upgraded, breaking through the limitations of early models and transforming the word segmentation task into a character-level sequence labeling problem—automatic segmentation is achieved by labeling word position information (such as B, M, E, S) for each character in the text, laying the foundation for the development of subsequent complex models.

1) CNN, RNN, LSTM, and BiLSTM

Convolutional Neural Networks (CNNs) excel at extracting local features from text, capturing local combination patterns between characters through convolutional kernels. They have good modeling capabilities for short-range word-formation information, but fail to efficiently capture long-sequence dependencies in text. Recurrent Neural Networks (RNNs) process text sequentially, preserving historical information and making them suitable for temporal tasks like Chinese word segmentation. However, they suffer from vanishing and exploding gradients, hindering their ability to effectively model long-range dependencies.

Long Short-Term Memory Networks (LSTMs) address the shortcomings of RNNs via gating mechanisms [16], preventing gradient vanishing while retaining long-term dependency information. In word segmentation tasks, they can stably learn contextual semantic relationships. Building upon this, Bidirectional Long Short-Term Memory Networks (BiLSTMs) encode text sequences simultaneously from left to right and right to left, fully utilizing contextual information. This aligns better with the requirement for considering both preceding and following context in Chinese word segmentation, making it the most basic and commonly used network structure in deep learning word segmentation.

2) BiLSTM-CRF classic architecture

BiLSTM-CRF is a classic and iconic architecture for deep learning Chinese word segmentation. It combines the contextual feature extraction capability of BiLSTM with the globally optimal sequence decoding capability of CRF [17], effectively solving the problems of lack of global constraints and unreasonable label prediction in a single BiLSTM model.

In this model, the BiLSTM layer is responsible for encoding the input character sequence and outputting the probability distribution of each label; the CRF layer introduces transition probability constraints between labels, performs global normalization and optimal path search on the entire sequence, and ensures that the output label sequence conforms to linguistic logic Compared to the method of independently predicting the label of each character, BiLSTM-CRF can eliminate unreasonable labeling results by utilizing global contextual constraints, achieving higher accuracy in word segmentation tasks with ambiguous fields and complex sentence structures, and has become the standard architecture for traditional deep learning word segmentation.

B. Word Segmentation Based on Pre-trained Models

The emergence of pre-trained language models has further improved the performance of Chinese word segmentation. This is achieved through pre-training on large-scale unlabeled corpora.

1) Applications of BERT, RoBERTa, etc. in word segmentation

BERT employs a bidirectional Transformer encoder structure, enabling deep bidirectional contextual representation learning and fundamentally transforming the feature extraction mode of word segmentation models [18]. When applied to Chinese word segmentation, the model can

directly learn character-level deep semantic representations, demonstrating a strong ability to understand complex linguistic phenomena such as ambiguity, ellipsis, and special sentence structures, significantly surpassing traditional neural network models in word segmentation accuracy.

RoBERTa optimizes the pre-training strategy based on BERT [19], removing the next sentence prediction task and adopting a training approach with larger batches and longer sequences, thereby further enhancing contextual representation capabilities. In Chinese word segmentation tasks, it can more stably capture word formation patterns between characters and has stronger adaptability to domain-specific and colloquial texts. Pre-trained models represented by BERT and RoBERTa have pushed Chinese word segmentation into a new stage where optimal results can be achieved with only simple fine-tuning, without the need for complex feature design.

C. Lightweight and Real-Time Word Segmentation Models

With the widespread application of word segmentation technology in low-resource scenarios such as mobile devices and embedded devices, lightweight and real-time word segmentation models have become an important research direction. These models, while maintaining segmentation accuracy, reduce the number of parameters and computational load through model compression, structural simplification, and knowledge distillation, thereby improving inference speed.

Common implementation methods include: pruning and quantizing large pre-trained models to construct miniaturized Transformer models; using lightweight BiLSTM structures to reduce the number of network layers and neurons; and utilizing knowledge distillation to transfer knowledge from large models to smaller models, significantly optimizing size and speed while maintaining high accuracy. Lightweight word segmentation models can meet the needs of real-time text processing and edge-side intelligent analysis, achieving a balance between segmentation accuracy and inference efficiency in word segmentation technology.

IV. Conclusion

This paper systematically reviews the technological evolution of Chinese word segmentation from traditional methods to deep learning paradigms, identifying three core problems: inconsistent general vocabulary standards, difficulty in ambiguous word segmentation, and poor recognition of out-of-vocabulary words. The study finds that dictionary-based methods are efficient but over-reliant on the vocabulary; statistical methods have stronger generalization capabilities but are limited by the corpus; rule-based methods can resolve ambiguities but are difficult to scale up. Deep learning, through automatic feature extraction and bidirectional context modeling, combined with the classic BiLSTM-CRF architecture and pre-trained models such as BERT, has become the mainstream solution to overcome these traditional challenges. Most mainstream word segmentation systems also adopt a fusion approach of traditional methods and deep learning.

This study only provides a theoretical overview through literature review and does not conduct quantitative experiments. Furthermore, it lacks sufficient discussion on domain-specific word segmentation. Future research could focus on deeply integrating large language models with traditional word segmentation algorithms to optimize the accuracy of ambiguous word and out-of-vocabulary word recognition; developing lightweight specialized models for fields such as medicine and law; and exploring real-time word segmentation technology on the edge to achieve a better balance between accuracy and efficiency, thereby providing stronger support for high-level tasks in Chinese natural language processing.

Conflict of Interest

The authors declare no conflict of interest.

Reference

[1] D. Jurafsky, J. H. Martin, *Speech and Language Processing*, Pearson, 2023.

[2] S. W. Yu, H. M. Duan, X. F. Zhu *et al.*, "Basic processing specifications of Peking University modern chinese corpus," *Journal of Chinese Information Processing*, vol. 16, 2002.

[3] Q. Liu, X. D. Wang, H. Liu, *et al.*, "HTRDP evaluations on Chinese information processing and intelligent human-machine interface," *Frontiers of Computer Science in China*, vol. 1, no. 1, pp. 58–93, 2007.

[4] M. Sun and H. Wang, *Chinese Lexical Analysis: Methods and Applications*, San Rafael: Morgan & Claypool Publishers, 2016.

[5] J. Devlin, M. W. Chang, K. Lee *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[6] M. Sun and J. Zou, "A survey of research on Chinese word segmentation," *Contemporary Linguistics*, vol. 3, no. 1, pp. 22–32, 2001.

[7] Y. Wu, G. Wei, H. Li, "A Chinese word segmentation algorithm based on n-gram model and machine learning," *Journal of Electronics & Information Technology*, vol. 23, no. 7, pp. 1148–1153, 2001.

[8] F. Zhang and X. Zhang, "Research on optimization of Chinese word segmentation based on hidden markov model," *Science of Surveying and Mapping*, no. 2, pp. 44–48, 2025.

[9] L. Jia, "Research on word segmentation technology based on maximum entropy model," Jinan: Shandong Normal University, 2007.

[10] J. Lafferty *et al.*, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th International Conf. on Machine Learning (ICML)*, 2001, pp. 282–289.

[11] J. Zhang, "Rule-based word segmentation method," *Computer and Modernization*, no. 4, pp. 18–20, 2005.

[12] W. Zhao, X. Dai, C. Yin *et al.*, "A Chinese word segmentation method combining rules and statistics," *Computer Application Research*, no. 3, pp. 23–25, 2004.

[13] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, pp. 504–507, 2006.

[14] Y. LeCun, L. Bottou, Y. Bengio *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, 1999.

[15] Z. L. Pei, R. N. A., M. Y. Jiang *et al.*, "a survey on text classification research based on convolutional neural networks," *Journal of Inner Mongolia Minzu University (Natural Sciences Edition)*, vol. 34, no. 3, pp. 1–8, 2019.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] Z. Huang, W. Xu, and K. Yu, "Bidirectional LSTM-CRF models for sequence tagging," arXiv preprint arXiv:1508.01991, 2015.

[18] J. Devlin, M. W. Chang, K. Lee *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[19] Y. Liu, M. Ott, N.Goyal *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

Copyright © 2026 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).