Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/37/20230464

Research advanced in Chinese word segmentation methods and challenges

Guancheng Du

Beijing No.18 High School, Beijin, 100005, China

shuoren100@163.com

Abstract. Chinese word segmentation refers to the process of dividing a sequence of Chinese characters into individual words. It constitutes a fundamental component of Chinese natural language processing. Due to the intricacies of the Chinese language, Chinese word segmentation has garnered significant attention from researchers. Based on a review of historical literature, segmentation methods can be broadly categorized into rule-based, statistical, semantic-based, and comprehension-based approaches. With the advancement of machine learning, neural networks have emerged as the mainstream algorithm for word segmentation. However, Chinese presents several unique challenges, leading to segmentation results that are less effective compared to morphological analysis in languages like English. Moreover, word segmentation faces new challenges such as dependency on the quality and scale of corpora, as well as domain-specific segmentation in diverse fields. Addressing these emerging challenges will undoubtedly become a focal point in future research endeavors in this field. This review provides a comprehensive summary of existing methods, discusses the current state of Chinese word segmentation, and outlines directions for addressing the evolving complexities in the field. As Chinese language processing continues to advance, finding robust solutions for accurate word segmentation remains a critical area of research.

Keywords: Chinese word segmentation, Natural Language Processing, neural networks, cross-domain word segmentation.

1. Introduction

The rapid advancement of artificial intelligence has propelled Natural Language Processing (NLP) tasks into pivotal roles across various scenarios, including human-computer interaction, personalized recommendations, intelligent search, and risk management. In comparison to English, the Chinese language presents challenges of ambiguous word boundaries and intricate grammatical structures, which hinder the direct processing capabilities of computers. Chinese Word Segmentation (CWS) addresses these challenges by employing diverse methods to achieve precise segmentation and part-of-speech tagging of text, making it a foundational step in NLP tasks. However, recent scholarly discourse has sparked debates over the necessity of CWS research.

In 2019, Li et al. [1] conducted comparative experiments involving four NLP benchmark tasks, revealing that deep learning-based word-level models exhibited susceptibility to overfitting due to issues like out-of-vocabulary words, data sparsity, and cross-domain complexities, thereby underperforming compared to character-level models. Nevertheless, this should not diminish the significance of

DOI: 10.54254/2755-2721/37/20230464

segmentation research. In 2020, Chen et al. [2] argued that the lack of word-level information in character-level models holds potential benefits for text matching tasks. Furthermore, in NLP tasks involving terminology, such as entity recognition, the effectiveness directly hinges on the foundational results of Chinese word segmentation. Yang et al. [3] even demonstrated a marked enhancement in entity recognition performance by incorporating part-of-speech results derived from word segmentation. This underscores the enduring significance of Chinese word segmentation research.

As traditional methods are gradually supplanted by deep learning techniques, becoming the mainstream approach for segmentation research, this paper delves into the landscape of CWS technology research over the past five years, both domestically and internationally. It provides an introduction, summary, and analysis of the current state of research involving traditional and deep learning methods, as well as the associated challenges. Furthermore, it explores the focal points and potential directions for the future development of CWS technology. By doing so, this paper aims to offer insights and guidance for subsequent research endeavors in the field.

2. Method

In light of the distinct characteristics inherent in Chinese text, the realm of Chinese Word Segmentation (CWS) has evolved diverse algorithms aimed at deciphering the intricacies of its linguistic structure [4]. These algorithms can be broadly categorized into four main approaches, each tailored to tackle the unique challenges posed by Chinese language processing. This section provides a comprehensive overview of these algorithmic categories while simultaneously delving into the pivotal challenges that pervade the landscape of Chinese word segmentation. By classifying these algorithms and illuminating the associated difficulties, we pave the way for an in-depth exploration of the methodologies employed in this crucial domain of Natural Language Processing.

2.1. Rule-based segmentation methods

Rule-Based Word Segmentation Methods, also referred to as mechanical segmentation methods or dictionary-based segmentation methods, are an early category of approaches employed in Chinese Word Segmentation (CWS). This methodology entails matching a given sequence of Chinese characters to entries within an extensive machine-readable lexicon, employing predefined strategies. When a matching entry is located within the lexicon, successful segmentation is achieved.

Key constituents of this method include the segmentation lexicon, the scanning sequence of the input text, and the matching principles [5]. Notably, the scanning sequence encompasses three primary directions: forward scanning, reverse scanning, and bidirectional scanning. Correspondingly, the matching principles encompass diverse strategies such as the maximum matching method, minimum matching method, word-by-word matching method, and optimal matching method.

(1) Maximum Matching Method (MM). This approach assumes that the longest word within the segmentation lexicon contains "i" characters [6]. It involves extracting the initial "i" characters from the current string in the text and seeking a corresponding entry within the lexicon. When such an entry is found, successful segmentation is achieved. If not, the last character is removed from the current segment, and the process iterates until a match is found, forming a coherent word.

(2) Reverse Maximum Matching Method (RMM). Similar to MM, this method commences segmentation from the end of the sentence, iteratively removing characters from the beginning of the sentence until a match is identified within the lexicon.

(3) Word-by-Word Traversal Method. In this technique, words within the lexicon are systematically searched from longest to shortest, character by character, through the entire input text. This exhaustive approach ensures comprehensive coverage, regardless of the lexicon's size or the input text's length [7].

(4) Boundary Marking Method. This approach involves identifying and incorporating both natural and non-natural boundaries for segmentation. Natural boundaries encompass punctuation marks, while non-natural boundaries involve prefixes or non-word elements. This preliminary step is followed by additional processing using methods like MM or RMM, constituting a preprocessing technique rather than a definitive segmentation method.

DOI: 10.54254/2755-2721/37/20230464

(5) Optimal Matching Method (OM). OM consists of forward and reverse variants [8]. It organizes lexicon entries based on their frequency, aiming to expedite retrieval and enhance efficiency. OM contributes to reducing the time complexity of segmentation while sustaining or even improving accuracy.

Despite the simplicity and ease of implementation associated with rule-based segmentation, these methods possess limitations. They exhibit relatively slow matching speeds, grapple with both intersectional and compositional ambiguity, lack standardized definitions for words, and yield different ambiguities across various dictionaries. Moreover, they lack the capacity for self-learning and adaptability, which restricts their ability to accommodate dynamic language use. Nevertheless, they serve as foundational approaches in the landscape of Chinese word segmentation, paving the way for subsequent advancements in more sophisticated methodologies.

2.2. Statistic-based segmentation methods

Statistic-Based Word Segmentation Methods leverage statistical machine learning models to uncover patterns of word segmentation, capitalizing on abundant pre-segmented text data to facilitate segmentation of previously unseen text [9]. The core concept revolves around the notion that words constitute stable combinations, and the higher the frequency of adjacent characters appearing together in context, the more likely they form a coherent word. Hence, the probability of neighboring character co-occurrences serves as a reliable indicator of word validity. By statistically analyzing the frequency of character combinations in training texts and calculating their mutual information, one can gauge the closeness of character associations. When this closeness exceeds a predefined threshold, the character group is deemed likely to form a word. This methodology is commonly referred to as dictionary-free segmentation.

A range of statistical models can be harnessed for this method, including N-gram models [10], Hidden Markov Models (HMM) [11], Maximum Entropy (ME) models, and Conditional Random Fields (CRF) [12]. These models capitalize on statistical patterns to predict word boundaries, offering varying degrees of contextual understanding.

In practical applications, this segmentation method frequently combines the use of segmentation lexicons for string matching with statistical methodologies for identifying new words. This amalgamation optimally marries the strengths of both approaches. On one hand, the efficiency of string matching enables rapid and effective segmentation. On the other hand, the incorporation of statistical techniques facilitates the recognition of novel terms based on their contextual significance, as well as the automatic resolution of ambiguities.

The potency of statistic-based segmentation lies in its adaptability to diverse domains and the capacity to integrate pre-existing knowledge while accommodating evolving linguistic nuances. By harnessing statistical insights from extensive corpora, these methods refine segmentation accuracy, rendering them valuable tools in the realm of Chinese word segmentation.

2.3. Semantic-based segmentation methods

Semantic-based word segmentation methods incorporate semantic analysis to process linguistic information more extensively within natural language, encompassing various techniques such as the Extended Transition Network Method, Knowledge-Based Semantic Analysis, Adjacent Constraints, Comprehensive Matching, Suffix Segmentation, Lexicon-Based Feature Method, Matrix Constraints, and Syntax Analysis.

Extended Transition Network Method. Rooted in the concept of finite-state machines, this method employs the Recurrent Transition Network (RTN) to enhance the capabilities of a finite-state machine, enabling recursive transitions. RTN incorporates terminal symbols (words in the language) or non-terminal symbols (part-of-speech categories) on its arcs. It can also invoke other sub-networks, serving as non-terminal symbols, to define word formations (such as word or character sequences). This hierarchical network structure facilitates interaction between segmentation and syntactic parsing stages, effectively resolving ambiguities in Chinese word segmentation.

DOI: 10.54254/2755-2721/37/20230464

Matrix Constraints. This approach involves establishing both a grammar constraint matrix and a semantic constraint matrix. Elements within these matrices indicate whether words with specific part-of-speech tags or semantic classes should be adjacent based on grammatical rules or logical coherence. During the segmentation process, the machine adheres to these matrices to guide segmentation outcomes.

These methods intrinsically aim to leverage semantic information to enhance segmentation accuracy and contextuality. By fusing linguistic and contextual insights, they bridge the gap between syntax and semantics, contributing to a more holistic understanding of the Chinese language structure. Incorporating semantic context into the segmentation process addresses challenges such as ambiguity resolution and contributes to the development of advanced segmentation techniques. These techniques underscore the evolving landscape of Chinese word segmentation, advancing beyond traditional rule-based and statistical methods to harness the intricacies of semantic relationships and linguistic context. By delving into the semantic layers of the language, these methods contribute to the refinement of Chinese word segmentation systems for enhanced accuracy and comprehension.

2.4. Understanding-based segmentation methods

Understanding-Based Word Segmentation Methods involve simulating human sentence comprehension processes within computers to achieve word recognition. The fundamental idea is to conduct syntactic and semantic analyses concurrently with word segmentation, utilizing syntactic and semantic information to address ambiguity. Typically, this approach comprises three main components: the segmentation subsystem, the syntax-semantics subsystem, and the control module. Under the coordination of the control module, the segmentation subsystem leverages syntactic and semantic information related to words and sentences to make judgments about segmentation ambiguities, simulating human sentence comprehension. This method necessitates substantial linguistic knowledge and information. Presently, understanding-based segmentation methods primarily encompass Expert Systems Segmentation and Neural Network Segmentation.

Expert Systems Segmentation. This approach segregates the knowledge required for segmentation, including common sense segmentation knowledge and heuristic knowledge for disambiguation (ambiguity disambiguation rules), from the reasoning engine that executes the segmentation process. This decoupling facilitates the independent maintenance and management of the knowledge base and reasoning engine. Expert systems segmentation possesses the capability to detect intersectional ambiguities and polysemous combination ambiguities, along with a degree of self-learning.

Neural Network Segmentation. This method emulates the parallel, distributed processing and numerical modeling capabilities of the human brain. It stores segmentation knowledge implicitly within neural network structures, modifying internal weights through self-learning and training to achieve accurate segmentation outcomes. Techniques such as Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and other neural network models are commonly employed in this context.

Integrated Neural Network and Expert System Segmentation. This methodology initiates segmentation using neural networks. When neural networks encounter new words for which they cannot provide precise segmentation, the expert system is activated for analysis and judgment. It relies on knowledge bases to infer and generate preliminary analyses, thereby initiating a learning mechanism to train the neural network. This approach effectively harnesses the strengths of both neural networks and expert systems, further enhancing segmentation efficiency.

Understanding-based segmentation methods exemplify the advanced stage of segmentation techniques, capitalizing on human-like language comprehension to disambiguate complex linguistic contexts. As technology advances, the integration of linguistic knowledge and machine learning capabilities continues to yield refined segmentation outcomes, contributing to the ongoing evolution of Chinese word segmentation methodologies.

DOI: 10.54254/2755-2721/37/20230464

2.5. Challenges in Chinese word segmentation

Chinese word segmentation encounters numerous distinct challenges that hinder its effectiveness when compared to morphological analysis in languages such as English. These difficulties primarily manifest as follows:

(1) Lack of Unified Word Definitions: The absence of a consistent and universally accepted definition for words in Chinese contributes to segmentation complexity. Varied interpretations of what constitutes a word in different contexts lead to ambiguity.

(2) Absence of Standard Segmentation Guidelines: Chinese segmentation lacks a widely acknowledged standard, further compounded by the absence of a unified word definition. This issue stems from the preceding challenge and creates uncertainties in segmentation practices.

(3) Ambiguities in Word Determination: The precise identification of word boundaries remains a persistent challenge. The intricacies of Chinese linguistic structure and context often lead to ambiguities that require sophisticated techniques for resolution.

(4) Lack of Natural Language Formal Models: Chinese word segmentation encounters difficulties in devising comprehensive and accurate natural language formal models. Unlike languages with explicit morphological rules, the character-based nature of Chinese compounds these challenges.

(5) Effective Utilization of Syntax and Semantic Knowledge: Efficiently harnessing and representing the syntactic and semantic knowledge necessary for segmentation poses a significant challenge. Integrating these linguistic dimensions into the segmentation process to enhance accuracy remains an ongoing endeavor.

(6) Semantic Understanding and Formalization: Converting semantic understanding into formal rules and representations presents inherent complexities. Capturing the nuanced meanings and relationships between characters and words requires advanced semantic processing techniques.

Addressing these challenges necessitates a multifaceted approach that blends linguistic insights, computational methodologies, and machine learning techniques. As Chinese word segmentation matures, researchers continue to explore innovative solutions to enhance segmentation accuracy, adaptability, and effectiveness in capturing the nuances of the Chinese language structure.

3. Experiment

3.1. Dataset

The current experiment employs the THUCNews dataset, which consists of Chinese news headlines and is selected as the experimental data. Each headline comprises approximately 20 characters, rendering it suitable for Chinese word segmentation tasks. Within this dataset, 10 distinct categories have been chosen, encompassing politics, finance, sports, technology, education, entertainment, gaming, society, stocks, and real estate. Each category comprises 12,000 data instances, of which 10,000 are allocated for training purposes, while the remaining 1,000 are equally divided between validation and testing sets. This comprehensive dataset allows for robust evaluation and analysis of the Chinese word segmentation task across various domains.

3.2. Setting of parameters

The present experiment is conducted using Python 3.7 as the development language, with the PyTorch framework version 1.6.0 serving as the development platform. The operating system utilized is Ubuntu 16.04, and the GPU model employed is the NVIDIA TITAN Xp. Throughout the experiment, a dropout rate of 0.5 is consistently applied, and the learning rate is uniformly set to 0.001. These settings ensure a standardized experimental environment and promote reproducibility in the conducted tasks.

3.3. Results and comparison

In order to contrast the classification performance of four different models, namely TextCNN, BiLSTM, FastText, and BERT, on Chinese short texts, an experimental validation was conducted. The comparison

DOI: 10.54254/2755-2721/37/20230464

was based on four evaluation metrics: accuracy, precision, recall, and F1 score. The results are presented in Table 1 as follows:

Table 1. Experimental results comparison.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
TextCNN	90.89	90.92	90.89	90.90
BiLSTM	90.79	90.78	90.79	90.78
FastText	91.86	91.90	91.86	91.88
BERT	94.33	94.36	94.33	94.34

From the experimental results, it can be deduced that the classification performance of the pre-trained BERT model surpasses others, with FastText following as the second-best performing model, while TextCNN and BiLSTM models demonstrate slightly lower performance. The BERT model exhibits higher values across all four evaluation metrics compared to TextCNN, BiLSTM, and FastText models. Furthermore, it notably enhances the comprehensive evaluation metric, F1 score, by 3.44%, 3.56%, and 2.46%, respectively, for the other three classification models.

FastText also presents superior classification results across the four evaluation metrics compared to TextCNN and BiLSTM models. It elevates the F1 score by 0.98% and 1.1%, respectively, for TextCNN and BiLSTM models. Additionally, the FastText model boasts a quicker classification speed than the other three models.

In summary, the experimental findings emphasize that the BERT model exhibits the most favorable classification outcomes, closely followed by FastText, while TextCNN and BiLSTM models, although competent, lag behind in terms of classification performance and speed.

4. Conclusion

The field of Chinese word segmentation is undergoing gradual improvements. Rule-based segmentation methods, relying on dictionary-based mechanical approaches, offer simplicity and speed in execution. However, their reliance on existing dictionaries for direct matching introduces limitations in terms of domain specificity. Additionally, they struggle with ambiguity and out-of-vocabulary word recognition. Statistical segmentation techniques, particularly supervised learning algorithms, have addressed ambiguity issues by framing segmentation as a sequence labeling problem. The introduction of neural networks has significantly elevated segmentation accuracy.

Currently, research in the field of word segmentation has reached a level of maturity where it serves as a foundational task that supports the majority of Natural Language Processing (NLP) endeavors. Nevertheless, there exists a discrepancy between segmentation performance and speed. Convolutional neural networks (CNNs), due to their computational speed advantage, have gained prominence in recent years within the NLP domain, yielding noteworthy results. Chinese word segmentation research should strike a balance between accuracy and speed. The incorporation of CNNs, with their inherent speed advantage, is expected to contribute to substantial improvements in segmentation. Ultimately, achieving a harmonious interplay between accuracy and efficiency is essential for the ongoing advancement of Chinese word segmentation techniques.

References

[1] LI, X., MENG, Y., SUN, X., et al. Is word segmentation necessary for deep learning of Chinese representations? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 359-401.

[2] CHEN, L., ZHAO, Y., LYU, B., et al. Neural graph matching networks for Chinese short text matching. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 6152-6158.

[3] YANG, J. X., DU, J. P., SHAO, Y. X., et al. Construction method of intellectual-property-oriented scientific and technological resources portrait. Journal of Software, 2022, 33(4): 1439-1450.

DOI: 10.54254/2755-2721/37/20230464

[4] Luo, R., Xu, J., Zhang, Y., Zhang, Z., Ren, X., & Sun, X. (2019). Pkuseg: A toolkit for multi-domain chinese word segmentation. arXiv preprint arXiv:1906.11455.
[5] Li, P., Luo, A., Liu, J., Wang, Y., Zhu, J., Deng, Y., & Zhang, J. (2020). Bidirectional gated recurrent unit neural network for Chinese address element segmentation. ISPRS International Journal of Geo-Information, 9(11), 635.
[6] Yan, X., Xiong, X., Cheng, X., Huang, Y., Zhu, H., & Hu, F. (2021). HMM-BiMM: Hidden Markov Model-based word segmentation via improved Bi-directional Maximal Matching algorithm. Computers & Electrical Engineering, 94, 107354.
[7] Brouwer, H., Delogu, F., Venhuizen, N. J., & Crocker, M. W. (2021). Neurobehavioral correlates of surprisal in language comprehension: A neurocomputational model. Frontiers in Psychology, 12, 615538.
[8] Tian, X., & Jia, W. (2022). Optimal matching method based on rare plants in opportunistic social networks. Journal of Computational Science, 64, 101875.
[9] Baomao, P., & Haoshan, S. (2009, August). Research on improved algorithm for Chinese word segmentation based on Markov chain. In 2009 Fifth International Conference on Information Assurance and Security (Vol. 1, pp. 236-238). IEEE.
[10] Novak, J. R., Minematsu, N., & Hirose, K. (2016). Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework. Natural Language Engineering, 22(6), 907-938.
[11] Mor, B., Garhwal, S., & Kumar, A. (2021). A systematic review of hidden Markov models and their applications. Archives of computational methods in engineering, 28, 1429-1448.
[12] Yuan, H., & Ji, S. (2020, January). Structpool: Structured graph pooling via conditional random fields. In Proceedings of the 8th International Conference on Learning Representations.