Middle Chinese Automatic Word Segmentation System Based on CRFs and Dictionary Information

Authors: Wang Xiaoyu, Bin Li

Date: 2017-11-08T00:00:00+00:00

Abstract

Abstract

【Objective】 To verify the impact of word segmentation consistency and corpus type on CRFs segmentation efficiency in medieval Chinese texts, and on this basis, further improve segmentation efficiency and reduce manual proofreading workload.

【Method】 Taking historical texts, Buddhist scriptures, and novel-type corpora from the medieval period as examples, this study addresses the automatic word segmentation problem in medieval Chinese by optimizing segmentation principles and employing a method that combines CRFs models with dictionaries to eliminate common inconsistencies in manual segmentation results for medieval Chinese; simultaneously, two features—character classification and dictionary information—are introduced into CRFs segmentation, and the most appropriate segmentation template for each feature is selected through comparative experiments.

【Results】 Experimental results show that the overall F-score of segmentation results reaches over 99% in closed testing and 89%-95% in comprehensive open testing.

【Limitations】 The research on segmentation inconsistency primarily focuses on two-character words, thus the recognition performance for words of three or more characters (multi-character words) is somewhat limited.

【Conclusion】 Given the effective improvement in segmentation consistency, character classification and dictionary tagging features can effectively improve the accuracy of CRFs-based word segmentation for medieval Chinese; meanwhile, the medieval Chinese word segmentation system proposed in this paper can be applied to multi-category Chinese corpora from the medieval period.

Full Text

Automatic Word Segmentation for Middle Ancient Chinese Based on CRFs and Dictionary Information

Wang Xiaoyu, Li Bin
(School of Chinese Language and Literature, Nanjing Normal University, Nanjing 210097)

Abstract

Objective: This study investigates how segmentation consistency and corpus type affect the efficiency of Conditional Random Fields (CRFs) for word segmentation in Middle Ancient Chinese (MAC), aiming to improve segmentation accuracy and reduce manual proofreading workload.

Methods: Using historical records, Buddhist scriptures, and novels from the MAC period as experimental corpora, we optimized segmentation principles and employed a hybrid approach combining CRFs with dictionary information to eliminate segmentation inconsistencies common in manual annotation. We introduced two features—character classification and dictionary marking—into the CRFs model and identified optimal segmentation templates through comparative experiments.

Results: Experimental results show that the overall F-score exceeded 99% in closed tests and reached 89%-95% in open tests.

Limitations: The consistency research primarily focused on two-character words, leaving room for improvement in recognizing words with three or more characters.

Conclusions: Character classification and dictionary marking features effectively improve CRFs segmentation accuracy for MAC when segmentation consistency is properly addressed. The proposed MAC segmentation system can serve multiple corpus types from this period.

Keywords: Conditional Random Fields Model; Segmentation Consistency; Middle Ancient Chinese; Automatic Word Segmentation

Introduction

The boundary between words and phrases in Chinese is often ambiguous, a phenomenon particularly pronounced in Middle Ancient Chinese. In Chinese historical linguistics, the MAC period spans from the late Eastern Han Dynasty to the Sui Dynasty, representing a transformative phase when Chinese vocabulary shifted from predominantly single-character words to double-character words. This transition produced numerous character combinations occupying a gray area between words and phrases. Some were undergoing lexicalization, while others represented temporary multi-character combinations used as words.

These ambiguous combinations make the word-phrase boundary especially unclear in MAC. During corpus construction, individual annotators' linguistic intuitions vary, and the lexicalization timeline and degree of these combinations cannot be fully quantified, inevitably causing segmentation difficulties. This not only leads to segmentation inconsistency in manual annotation but also severely constrains the improvement of machine segmentation accuracy and consistency when using manually segmented data as training corpora. Word segmentation is fundamental to Chinese corpus construction and critically impacts subsequent annotation and semantic analysis tasks.

While MAC corpora are not as extensive as modern Chinese, they still comprise hundreds of millions of characters, including numerous historical records, Buddhist scriptures, folk literature, and miscellaneous writings [**?**]. Computer processing of MAC inevitably requires automatic word segmentation. However, research on ancient Chinese information processing remains limited, with even fewer studies focusing specifically on MAC segmentation. Wang Jialing [**?**] conducted automatic segmentation research on the *Book of Han*, establishing segmentation standards and achieving 94.4% F-score using CRFs with character classification and archaic phonological features. However, this study used only one book, which cannot represent the full scope of MAC corpora. Moreover, MAC materials exhibit significant variation across categories—not only between historical records, Buddhist scriptures, and miscellaneous writings but also within categories, such as differences between translated scriptures and biographies of monks. This substantially reduces the scalability of such research.

Wang Xiaoyu et al. [**?**] sampled 280,000 characters from representative MAC corpora, quantifying segmentation errors, inconsistencies, and combinatory ambiguities. They analyzed causes and categories of segmentation inconsistency and proposed solutions, covering Buddhist scriptures, historical records, and novels to reveal the general landscape of MAC manual segmentation and identify specific problems, laying groundwork for further research.

Building on these studies, this paper addresses MAC segmentation issues theoretically and practically. First, we formulated and optimized segmentation principles for strings prone to inconsistency, calibrating manual annotations to minimize errors and inconsistencies. Then, using the refined corpus as training data, we introduced character classification and dictionary marking features, testing multiple feature templates through comparative experiments to select the most effective ones. Finally, we designed two comparative experiments to verify how corpus type and segmentation consistency affect CRFs performance. These results directly serve MAC corpus construction.

2.1 Segmentation Principles

Although the word-phrase boundary in Chinese is often unclear, this does not hinder human language comprehension. Similarly, segmentation need not obsess over linguistic demarcation—so long as the system is applicable and granularity

appropriate, computers can correctly process linguistic units. This concept also applies to dictionary compilation, where stable phrases and idioms are included alongside words without forcibly distinguishing them. Consequently, the natural language processing field introduced the concept of "segmentation units"—basic units with definite semantic and grammatical functions [**?**].

MAC vocabulary contains many word-phrase ambiguous strings that cause segmentation inconsistency and severely impact corpus quality. Clarifying these boundaries is key to improving MAC corpus construction. The "Middle Ancient Chinese Lexicon" [**?**] (hereafter "the lexicon") represents a major achievement of the MAC corpus project, incorporating entries from *The Great Chinese Dictionary* [**?**], *Wei-Jin Southern and Northern Dynasties Words and Expressions* [**?**], *Middle Ancient Function Words and Expressions* [**?**], *Buddhist Dictionary* [**?**], and *Sutras Words Explanations* [**?**], with over 540,000 entries covering most MAC vocabulary.

Based on this lexicon and to balance uniformity, referentiality, and semantic completeness, we follow three principles for ambiguous strings:

1. **Inclusiveness Principle**: For word-phrase ambiguous strings that do not affect semantic understanding, prioritize combination over separation.
2. **Dictionary Principle**: Following Huang Juren's concept of segmentation reliability [**?**], any semantic unit included in the lexicon must be combined.
3. **Semantic Opacity Principle**: Semantic opacity includes four cases: new meanings through metaphor/metonymy, referential meaning shift, meaning loss in components, and part-of-speech change. Any unit meeting these criteria must be combined.

These optimized principles guide manual segmentation to produce training corpora. However, since training data comes from multiple annotators, inconsistencies persist [**?**]. Therefore, we further refined the training corpus to eliminate segmentation inconsistencies and designed comparative experiments using pre- and post-refinement data to verify consistency' s impact.

2.2 Experimental Corpus

Historical records and Buddhist scriptures constitute the most abundant MAC literature, supplemented by smaller amounts of folk literature (e.g., novels) and miscellaneous writings. We selected these two primary categories plus some novels as experimental corpora, detailed in Table 1 .

All corpora in Table 1 were preliminarily manually segmented and annotated. Following Section 2.1 principles, we proofread these results to minimize inconsistencies, creating pre- and post-consistency refinement versions. The post-proofreading test corpus served as the gold standard.

Three experimental setups used these materials: 1. For feature template selection, all manually segmented training corpora from Table 1 were used, with 1,000 characters selected from each of four test corpora as the combined test set.

2. For consistency impact comparison, pre- and post-refinement corpora served as training data, with test corpora as shown in Table 1. 3. For corpus mixture impact comparison, historical records, Buddhist scriptures, and the complete corpus were used separately as training data, with test corpora as shown in Table 1.

3 Model and Feature Template Selection

Statistical segmentation can be viewed as a text sequence classification problem, where the core task is identifying word boundaries—start, middle, and end positions. Segmentation models calculate the joint probability distribution of annotation sequences given observation sequences [**?**] to predict the most likely output. Among mainstream models like HMM, MEMM, and CRFs, CRFs demonstrate superior performance by overcoming HMM' s strong independence assumptions and MEMM' s label bias problem [**?**]. Therefore, we selected CRFs as our experimental model, adding different features and templates to identify the most effective combination.

3.1 Feature Selection

Applying CRFs to segmentation involves mining positional knowledge of characters in word formation from training corpora [**?**]. Features are core to CRFs segmentation, directly affecting accuracy. However, more features are not always better, as they increase computational load and risk interference from redundant data. We selected character classification and dictionary marking as features:

1. **Character Classification**: This coarse-grained classification of characters significantly improves segmentation accuracy for both modern [**?**] and ancient Chinese [**?**]. Our classification set is: T1 = {HZ, Punc, SenPunc, CNum, CCNum, D, X}, corresponding to Chinese characters, regular punctuation, sentence-ending punctuation, numerals, Heavenly Stems/Earthly Branches, the ordinal marker character, and unrecognized characters.

Since MAC uses Chinese characters for numbers that cannot be exhaustively enumerated, we created special categories for numerals and Heavenly Stems/Earthly Branches, which follow regular patterns. MAC texts also use traditional characters with variant and obscure forms, handled by the "unrecognized character" category. Both character classification and dictionary marking use automatic annotation.

2. **Dictionary Marking**: This dynamically marks characters' dictionary combination status. CRFs models rely on word frequency and contextual information, struggling with low-frequency words and showing strong corpus dependence. Two remedies exist: combining statistical methods with rules/dictionaries [**?**], or using large homogeneous training corpora.

However, MAC corpora are limited in size and heterogeneous, making the second approach impractical.

Therefore, we introduced dynamic dictionary marking based on the lexicon, using annotation set: T2 = {B, M, E, S, W, T, H, F}, representing word-beginning, word-middle, word-ending, single-character, punctuation, characters belonging to two words, three words, and four+ words respectively.

Using the opening of the *Sutra of One Hundred Parables*—"Thus have I heard: At one time the Buddha was staying in Rajagaha"—the marking process matches all possible segmentations against the lexicon, identifying the segments: "Thus have I heard," "thus," "at one time," "Rajagaha," and "Rajagaha city." The characters are then classified based on their positions in these segments: - Characters appearing only at word beginnings (e.g., the first characters of "Thus have I heard" and "one") are marked "B" - Characters appearing in two words (e.g., characters that appear in both "Thus have I heard" and "thus") are marked "T" - Characters appearing only at word endings (e.g., the final characters of "time" and "city") are marked "E" - Characters that function as single-character words (e.g., "Buddha" and "dwelt") are marked "S" - Punctuation is marked "W"

Based on character classification, dictionary marking, and proofread manual segmentation as gold standard, we obtained the standard CRFs corpus format shown in Table 2 .

3.2 Feature Template Comparison Experiments

In CRFs, feature templates effectively extract word boundary information, directly impacting segmentation quality. We added character classification and dictionary marking features sequentially, testing different templates for each. Results appear in Table 3 .

Using only literal information, two-character word F-score negatively correlated with overall F-score. The 1W+2C template ($\pm$1 character window, two-character co-occurrence) achieved the highest overall F-score, though with the lowest two-character word F-score, maintaining stable precision/recall ratios. Other templates over-segmented two-character words, increasing error rates.

Character classification effectively distinguishes characters, punctuation, and numerals, particularly improving segmentation of numeral-based units. Adding this feature improved overall F-score by ~6% and two-character word F-score. Templates 2C and 1W+2C correctly segmented three+ character numerals. The optimal character classification template is 2C.

Dictionary marking, as an authoritative resource, substantially boosted F-scores, reversing the negative correlation between two-character and overall F-scores. The best dictionary marking template is clearly 1W+2C, improving overall F-score by 39.91% over literal-only baseline.

In summary (bolded in Table 3), we selected: - Literal information + dictionary marking: 1W+2C template - Character classification: 2C template

The combined template is: Template-all = (2C) literal information + (1W+2C) character classification + (1W+2C) dictionary marking

4.1 Experimental Design

Theoretically, more uniform training corpora yield more regular segmentation patterns. However, MAC historical records and Buddhist scriptures exhibit significant linguistic and lexical differences, which may affect CRFs performance. We designed two comparative experiments:

Experiment 1: Consistency Impact - Using pre- and post-consistency refinement corpora as training data to examine how segmentation consistency affects CRFs results.

Experiment 2: Corpus Mixture Impact - Using separately and combined historical records and Buddhist scriptures as training data to examine how corpus heterogeneity affects CRFs results.

4.2 Evaluation Metrics

Using proofread manual segmentation as the gold standard, we evaluate using precision (P), recall (R), and F-score. The formulas are:

$$P = \frac{RW}{AW} \quad R = \frac{RW}{SW} \quad F = \frac{P \times R \times 2}{P + R}$$

Where RW = correctly segmented words by CRFs, AW = total words segmented by CRFs, and SW = total words in gold standard. P, R, and F range between 0 and 1, with values closer to 1 indicating better performance. P and R evaluate different aspects, while F reflects comprehensive performance.

5.1 Closed Test

Using consistency-refined training corpora, we trained CRFs models on pre- and post-refinement data and conducted closed tests to verify consistency' s impact. Results appear in Table 4 .

F-score is the primary metric. As shown in Table 4 (bolded), we conclude: 1. Post-consistency refinement enables more accurate boundary detection through context, character classification, and dictionary marking, significantly improving overall F-score: +15.70% for historical records and +12.38% for Buddhist scriptures. 2. Longer words show more significant improvement from consistency refinement. As word length increases, F-score gains grow more pronounced because inconsistencies primarily affect multi-character words, and dictionary marking strengthens these features. 3. Multi-character word F-score improved

by 58-67%, far exceeding overall and two-character word gains, because longer words appear less frequently and are more vulnerable to inconsistency interference. 4. Consistency refinement' s impact is slightly weaker for Buddhist scriptures due to their specific characteristics as translated texts with more linguistic variations, making regular pattern extraction more difficult.

Using post-refinement corpora, we tested historical records, Buddhist scriptures, and combined corpora separately to examine mixture impact. Results appear in Table 5 .

Table 5 shows that mixing different corpus types generally decreases F-score. However, the decline is minimal: distinguishing corpus types improved overall F-score by less than 0.3%. While MAC corpus categories differ, this variation' s impact on CRFs is much smaller than that of segmentation inconsistency.

5.2 Open Test

Following Table 1, we conducted open tests using pre- and post-consistency refinement training corpora to further verify consistency' s impact. Results appear in Table 6 .

Table 6 confirms closed test conclusions with one difference: while consistency refinement' s improvement is less dramatic than in closed tests, open test overall F-score still increased significantly: +6.98% for historical records and +4.13% for Buddhist scriptures.

Using post-refinement corpora, we tested historical and Buddhist scriptures separately versus combined to examine mixture impact in open tests. Results appear in Table 7 .

Combined with Table 5, Table 7 reveals: 1. In open tests without distinguishing corpus types, F-score increased, contrasting with closed tests where it decreased. 2. Compared to closed tests, open test overall F-score changes were larger but remained under 1%, far smaller and less stable than consistency' s impact.

These findings suggest that corpus differences affect human intuition, exacerbating inconsistency, so fine-grained classification improves results. However, after optimizing segmentation standards and reducing inconsistency, inter-category lexical differences prove smaller than expected. Fine-grained classification reduces training data per category, potentially decreasing efficiency.

Overall, fine-grained corpus classification contributes modestly to CRFs segmentation, but with limited MAC data and uneven category distribution, segmentation reduces per-experiment training data and may lower efficiency. Segmentation consistency is a crucial quality metric that substantially impacts overall results.

Conclusion

This study addresses MAC segmentation challenges by establishing principles, optimizing the process, and minimizing inconsistency through manual proofreading. By introducing character classification and dictionary marking features and selecting optimal templates through CRFs comparative experiments, we achieved a dictionary-statistics hybrid approach. Experiments demonstrate these features effectively utilize character combination patterns and existing lexicon information, improving overall F-score by 5% and 35% respectively. Consistency refinement significantly impacts automatic segmentation: closed test overall F-score improved by over 10%, open test by 3-7%. Corpus type distinction has minimal impact (<1%), but given limited MAC data, fine-grained classification reduces training data and may decrease efficiency. Therefore, with proper consistency handling, MAC automatic segmentation need not treat different corpus types separately.

Our consistency standards currently target two-character words. While multi-character word performance improved significantly, open test results remain around 60%—not yet ideal. Future work will: 1. Investigate multi-character word inconsistency and establish corresponding standards 2. Enhance the lexicon' s coverage of MAC segmentation units 3. Add longest-match marking to dictionary features to improve low-frequency multi-character word recognition

References

[1] Hua Zhenhong. Some Problems in the Deep Processing of the Medieval Chinese Corpus Construction[J]. Journal of Southwest University: Social Science Edition, 2014, 40(3): 136-142.

[2] Wang Jialing. Medieval Chinese Automatic Word Segmentation Using the "Book of Han" as an Example[D]. Nanjing: Nanjing Normal University, 2014.

[3] Wang Xiaoyu, Dong Zhiqiao. Investigation of the Causes of Inconsistency in Medieval Chinese Word Segmentation[J]. The Collected Papers of the Chinese History Study, 2015, 19: 20-33.

[4] GB-T13715-1992. Contemporary Chinese Language Word Segmentation Specification for Information Processing[S]. Beijing: China Standard Press, 1993.

[5] Luo Zhufeng, et al. The Great Chinese Dictionary[M]. Shanghai: Shanghai Lexicographical Publishing House, 2011.

[6] Cai Jinghao. Explanations of Words and Expressions from the Wei, Jin, and Southern and Northern Dynasties[M]. Nanjing: Jiangsu Ancient Books Publishing House, 1990.

[7] Dong Zhiqiao, Cai Jinghao. Explanations of Function Words and Grammar in Medieval Chinese[M]. Changchun: Jilin Education Publishing House, 1994.

[8] Ding Fubao. Buddhist Dictionary[M]. Beijing: China Bookstore Publishing House, 2011.

[9] Li Weiqi, Jiang Jicheng. Compilation and Explanation of Buddhist Scripture Terms[M]. Changsha: Hunan Normal University Publishing House, 2004.

[10] Huang Juren, Chen Kejian, Chen Fengyi, et al. Design Philosophy and Content of the "Chinese Word Segmentation Standard for Information Processing" [J]. Journal of Applied Linguistics, 1997(1): 94-102.

[11] Huang Changning, Zhao Hai. A Decade Review of Chinese Word Segmentation[J]. Journal of Chinese Information Processing, 2007, 21(3): 8-19.

[12] Wu Qiong, Huang Degen. Chinese Temporal Expression Recognition Based on Conditional Random Fields and Temporal Lexicon[J]. Journal of Chinese Information Processing, 2014, 28(6): 169-174.

[13] Duan Yufeng, Zhu Wenjing, Chen Qiao, et al. Research on Out-of-Vocabulary Identification Combining Conditional Random Fields with Domain Ontology Element Sets[J]. New Technology of Library and Information Service, 2015(4): 41-49.

[14] Xiu Chi. Research and Implementation of Chinese Word Segmentation Methods Adapted to Different Domains[D]. Beijing: Beijing University of Technology, 2013.

[15] Song Yan, Cai Dongfeng, Zhang Guiping, et al. A Chinese Word Segmentation Method Based on Character-Word Joint Decoding[J]. Journal of Software, 2009, 20(9): 2366-2375.

[16] Shi Min, Li Bin, Chen Xiaohe. CRF-Based Research on Integrated Word Segmentation and POS Tagging for Pre-Qin Chinese[J]. Journal of Chinese Information Processing, 2010, 24(2): 39-45.

[17] Zhao H, Kit C Y. An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework[C]//Proceedings of IJCNLP 2008, Hyderabad, India. 2008: 9-16.

Author Contributions

Li Bin: Conceived research idea, implemented software, revised manuscript quality.
Wang Xiaoyu: Conducted experiments, acquired and analyzed data, drafted and revised manuscript.

Conflict of Interest

All authors declare no conflict of interest.

Supporting Data

Supporting data is available in the journal's online version at http://www.infotech.ac.cn.

[1] Wang Xiaoyu, Li Bin. Medieval Chinese CRFs Word Segmentation Test Data Combined with Dictionary.

Note: Figure translations are in progress. See original paper for figures.