applied sciences

MDPI

Article

SegFusion: A Lattice-Based Dynamic Ensemble Framework for Chinese Word Segmentation with Unsupervised Statistical Features

Chengfeng Wen [1,2] and Jiqiu Deng [1,2,*]

[1] School of Geosciences and Info-Physics, Central South University, Changsha 410083, China
[2] Key Laboratory of Metallogenic Prediction of Nonferrous Metals and Geological Environment Monitoring of Ministry of Education, Central South University, Changsha 410083, China
* Correspondence: csugis@csu.edu.cn

Abstract

Although existing Chinese word segmentation systems have achieved substantial progress on standard benchmarks, prediction disagreements among heterogeneous models remain prevalent when processing texts containing complex ambiguities and out-of-vocabulary words, and traditional static ensemble methods such as majority voting often fail to make reliable decisions in low-consensus scenarios. To address this issue, this paper proposes SegFusion, a stacked heterogeneous ensemble framework for Chinese word segmentation based on word lattice re-scoring. The framework first constructs a candidate word lattice to consolidate diverse outputs from heterogeneous segmenters into a unified lattice representation, and then incorporates unsupervised statistical features, including mutual information and branching entropy, as external discriminative evidence to perform dynamic arbitration at the word level, followed by global decoding to obtain the optimal segmentation path. Experimental results on multiple standard datasets demonstrate that SegFusion consistently outperforms individual models and mainstream ensemble baselines in terms of overall segmentation performance and out-of-vocabulary (OOV) recall. In particular, on the MSR dataset with severe ambiguity, SegFusion achieves improvements of 3.71% in F1 score and 4.10% in OOV recall. Further fine-grained analysis shows that the introduction of unsupervised statistical features effectively mitigates model consistency bias in low-support scenarios. These results indicate that integrating language statistical priors independent of training data into the ensemble arbitration stage is an effective way to enhance the robustness and consistency of Chinese word segmentation systems.

Keywords: Chinese word segmentation; ensemble learning; word lattice; out-of-vocabulary words; unsupervised statistical features; meta-arbitration

Academic Editor: Christos Bouras

Received: 30 January 2026
Revised: 27 February 2026
Accepted: 2 March 2026
Published: 4 March 2026

1. Introduction

Chinese word segmentation (CWS) is a fundamental task in Chinese natural language processing, and its performance directly affects the overall effectiveness of downstream applications such as named entity recognition, relation extraction, syntactic parsing, and machine translation [1–5]. In recent years, with the development of deep neural networks and pretrained language models, segmentation systems have achieved relatively high performance on standard evaluation benchmarks. However, in texts containing complex ambiguities, domain-specific named entities, and out-of-vocabulary words (OOV), disagreements among different segmentation systems remain widespread. Such disagreements are

often concentrated at a small number of critical positions within a sentence, yet they can produce cascaded and amplified negative effects on downstream semantic understanding [6]. Moreover, similar word boundary ambiguities are also observed in many languages without explicit word delimiters, such as Japanese, Thai, and Vietnamese, indicating that the challenges addressed in this study extend beyond Chinese.

Taking the string "东方红三号" as an example, this expression usually refers to a complete named entity in real contexts, but different segmentation models may produce results such as "东方红/三号", "东方红/三/号", or "东方红三号". These segmentations are all locally plausible at the boundary level, but they correspond to different entity hypotheses at the word-level semantic structure. This phenomenon indicates that complex ambiguous structures exhibit substantial diversity, and different segmentation models demonstrate selective coverage in their recognition capabilities. As a result, a single model is unlikely to maintain stable performance across all ambiguity scenarios [7].

From a methodological perspective, existing CWS techniques can be broadly categorized into three classes: dictionary- and rule-based methods, statistical sequence labeling approaches, and deep neural network-based models [8]. Due to differences in modeling assumptions and inductive biases, these models often exhibit complementary behaviors in complex ambiguity and cross-domain scenarios. Based on this observation, ensemble learning has been introduced to fuse the predictions of multiple segmenters, yielding certain improvements in overall performance [9,10]. Nevertheless, existing segmentation ensemble methods still suffer from notable limitations. Most approaches implicitly adopt a majority-consensus assumption and primarily perform static fusion at the character level. In low-consensus scenarios, especially for OOV words and long-word structures, correct segmentations are often supported by only a small number of models. Simple voting or static weighting schemes are therefore prone to amplifying the erroneous biases of dominant models [11]. Even when learnable stacked arbitration mechanisms are employed, their ability to correct consistency errors remains limited if the discriminative evidence is derived solely from the outputs of the base segmenters themselves.

Based on the above analysis, this paper explicitly argues that the performance bottleneck of segmentation ensembles in complex ambiguity and OOV scenarios arises from two structural limitations at the fusion stage. First, there is a lack of global structural context under model disagreement. When heterogeneous segmenters exhibit pronounced disagreements at critical boundaries, traditional fusion strategies that rely solely on base model outputs or character-level local information fail to capture word-level structural consistency. As a result, long words and compound structures cannot be reliably assessed in a global context. Second, there is a systematic suppression of minority but valid candidates. In low-consensus scenarios, correct segmentations are often supported by only a small subset of models. Majority-based voting or static fusion strategies therefore tend to systematically suppress these minority but semantically valid candidates, leading to persistent errors in ambiguous and OOV cases.

To address these structural limitations, this paper proposes a dynamic meta-arbitration framework that performs global modeling over word-level candidate structures and incorporates unsupervised statistical cues. Accordingly, this study aims to systematically investigate the effectiveness of the proposed SegFusion framework through the following research tasks:

(1) **Experimental Protocol Establishment:** to identify appropriate benchmark datasets and evaluation metrics, and to select representative baseline methods for fair and consistent evaluation;
(2) **Effectiveness of Heterogeneous Integration:** to examine the effectiveness of cross-paradigm integration across diverse benchmark datasets;

(3) **Comparison with Advanced Systems:** to compare the proposed framework with state-of-the-art (SOTA) segmentation and ensemble approaches;
(4) **Upper Bound and Complementarity Analysis:** to analyze the performance upper bound and the complementarity among different segmenters;
(5) **Ablation Study:** to assess the contribution of individual components within the proposed framework;
(6) **Fine-Grained Analysis:** to conduct detailed analyses with respect to OOV words and word-length distributions.

The main contributions of this paper are summarized as follows:

- A dynamic meta-arbitration ensemble framework based on word lattice re-scoring is proposed, which achieves unified modeling of heterogeneous segmenter outputs in the word-level structural space without requiring retraining of the underlying base segmentation models.
- Unsupervised statistical priors are introduced: mutual information and branching entropy are leveraged as external discriminative evidence, effectively addressing the failure of purely supervised ensembles in low-consensus and OOV scenarios.
- Experimental results on multiple standard datasets, including ICWB-2, demonstrate that the proposed method outperforms single models and static ensemble baselines in terms of both overall F1 score and OOV recall.

The remainder of this paper is organized as follows. Section 2 reviews related work on Chinese word segmentation and ensemble learning. Section 3 presents the overall architecture of the proposed SegFusion framework, including candidate generation, word lattice construction, dynamic meta-arbitration, and the incorporation of unsupervised statistical features. Section 4 reports the experimental setup and evaluation results corresponding to the research tasks outlined above. Section 5 provides an in-depth discussion of the experimental findings and analyzes the strengths and limitations of the proposed framework. Finally, Section 6 concludes the paper and outlines directions for future research.

2. Related Work

2.1. Overview of Chinese Word Segmentation

Chinese word segmentation (CWS) has evolved through several major paradigms. Early approaches primarily relied on manually constructed lexicons and heuristic matching rules [12]. With the increasing availability of annotated corpora, statistical learning approaches gradually became dominant, formulating CWS as a character-level sequence labeling task using models like hidden Markov models (HMMs), maximum entropy Markov models (MEMMs), and conditional random fields (CRFs) [13–15]. Subsequently, deep neural networks, particularly BiLSTM-CRF architectures, significantly enhanced the ability to capture long-range contextual dependencies and became widely adopted baselines [16]. Building upon this, pre-trained language models such as BERT further improved character-level semantic representations [17–21].

Recently, the paradigm has rapidly expanded to include Large Language Models (LLMs), which have demonstrated remarkable zero-shot capabilities in handling severe out-of-vocabulary (OOV) terms and highly heterogeneous texts. State-of-the-art studies propose a "comprehend first, segment later" philosophy to push the limits of unsupervised word segmentation [22], and leverage LLMs with knowledge-enhanced prompting or self-training to segment complex domain-specific terminology [23,24]. However, directly deploying these massive generative models for foundational character-level tokenization introduces prohibitive inference latency and computational overhead, limiting their viability in real-time or resource-constrained applications.

Meanwhile, to address the limitations of pure character-based models in exploiting lexical boundaries, structured modeling approaches such as lattice-based methods have been proposed [25–28]. By explicitly introducing multi-granularity word candidates, lattice structures enhance global structural consistency. The necessity of explicitly modeling word-level structural dependencies has been reaffirmed by recent advancements, such as integrating structural embeddings into LLM in-context learning to further resolve boundary ambiguities [29].

Despite these continuous improvements, different segmentation paradigms still exhibit strong complementarity in handling OOV words and ambiguous structures. This suggests that there remains substantial room for ensemble-based fusion strategies to further enhance robustness under model disagreement.

2.2. Ensemble Learning for Chinese Word Segmentation

Ensemble learning aims to improve system robustness and generalization by combining the predictions of multiple base learners. In the context of Chinese word segmentation, existing studies differ substantially in both fusion strategies and decision levels. Broadly speaking, one line of work performs static fusion of model outputs at the character or boundary level using fixed rules, while another line explores learnable arbitration mechanisms to re-evaluate predictions from multiple segmenters.

Early ensemble approaches for CWS were primarily built upon character-level sequence labeling frameworks such as BMES or BIO. In this setting, outputs from different segmentation systems are mapped to unified character-level tags and combined through majority voting. Representative work in the SIGHAN shared task integrated dictionary matching with multiple CRF-based sub-models, achieving improved segmentation accuracy through character-level voting and post-processing strategies [30]. Subsequently, analogy-based mechanisms were further combined with majority voting, providing additional empirical evidence for the effectiveness of this paradigm [31].

Beyond static output-layer fusion, some studies introduced data perturbation strategies during training to enhance ensemble stability. For example, bootstrap resampling was employed to train multiple segmenters at different granularities, and their outputs were fused through voting, demonstrating the robustness benefits of bagging-style ensembles in CWS [32].

Overall, static ensemble methods typically assign fixed or approximately equal weights to base models, making it difficult to capture variations in model reliability across different contexts or domains. Moreover, their decisions are largely confined to character labels or local boundary positions, resulting in weak consistency constraints at the word level. As a consequence, such approaches often encounter performance bottlenecks in scenarios involving overlapping ambiguities and long-word structures.

In addition to fixed-rule-based voting methods, a limited number of studies have explored learnable re-ranking or re-discrimination mechanisms for CWS ensembles. These approaches employ multi-stage architectures or secondary classifiers to correct the outputs of multiple segmentation systems, including cross-domain re-discrimination [33], two-stage decoding frameworks with intermediate structural constraints [34], and deep stacking strategies for neural fusion [35]. Compared with static voting, such methods partially relax the rigid majority-consensus assumption and enable more flexible decision-making.

However, research along this direction remains relatively limited and is often tailored to specific problem settings, such as domain adaptation, low-resource transfer, or joint modeling, rather than providing a general-purpose ensemble framework for Chinese word segmentation. More critically, the discriminative evidence used in existing learnable arbitration methods is largely derived from the outputs or internal representations of the base

segmenters themselves. When multiple models simultaneously produce disagreements or even consistent biases on OOV words or complex ambiguous structures, the arbitration stage lacks external references independent of supervised training distributions, thereby constraining its ability to correct such errors.

Against this backdrop, unsupervised statistical features derived from large-scale unlabeled corpora offer a promising avenue for extending the discriminative basis of the arbitration stage. By characterizing string cohesion and boundary freedom, these features provide language-intrinsic statistical consistency cues that are independent of specific segmentation models. Nevertheless, in existing CWS research, unsupervised statistical features are typically employed as heuristic rules or standalone new-word discovery modules, and are rarely integrated in a systematic manner into learnable ensemble arbitration frameworks.

Overall, prior studies have demonstrated the effectiveness of diverse modeling paradigms and ensemble strategies for Chinese word segmentation. However, existing methods still exhibit clear limitations in handling low-consensus ambiguities and out-of-vocabulary words, indicating that there remains substantial room for further investigation from the perspective of word-level structural modeling and the incorporation of external unsupervised statistical evidence.

3. Methodology

3.1. Framework Overview

In this section, we present the proposed SegFusion framework. While the base segmenters and the mathematical formulations of the statistical measures employed in this study follow established paradigms, the SegFusion framework introduces a fundamentally different ensemble formulation for Chinese word segmentation. Specifically, SegFusion performs dynamic meta-arbitration directly at the result level by re-scoring candidate words within a unified word lattice, enabling global structural modeling beyond traditional character-level voting or static fusion. Moreover, unsupervised statistical evidence is systematically integrated as external and model-independent signals to support the arbitration process, thereby alleviating segmentation conflicts in low-consensus and out-of-vocabulary (OOV) scenarios. Accordingly, although SegFusion builds upon established components, the overall ensemble methodology and arbitration formulation are proposed by the authors in this study.

The overall workflow is illustrated in Figure 1. Given an input sentence, the system first employs multiple base segmenters with different modeling paradigms to generate candidate segmentation results, which are then consolidated into a unified candidate space. Subsequently, each candidate word is represented as a feature vector and assigned a confidence score by a learnable meta-model, indicating its likelihood of being a correct segmentation unit. Finally, all candidate words are organized into a weighted word lattice, and global decoding is performed to obtain the final segmentation output. Unlike traditional ensemble methods based on character-level label fusion, the proposed framework operates directly at the candidate word level, enabling multi-source information to participate in decision-making in a more flexible and interpretable manner.

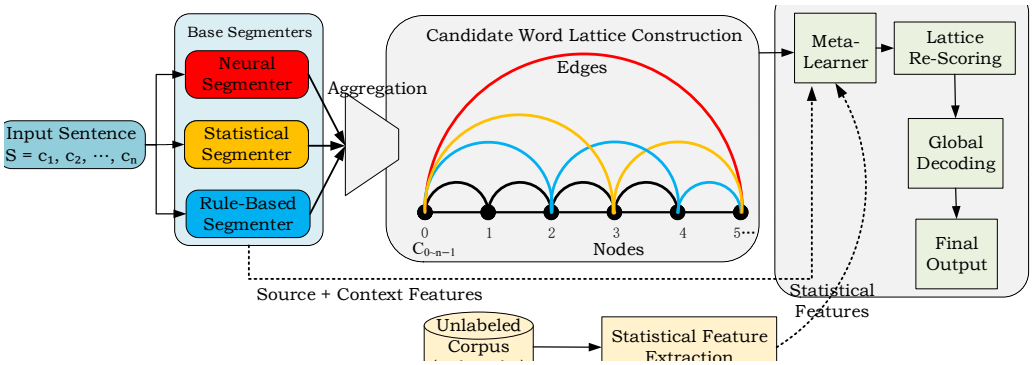

Figure 1. Architecture of the proposed framework. The red, orange, and blue colors represent the components of the Neural, Statistical, and Rule-based segmenters (shown on the left) and their corresponding lattice edges, respectively. The black edges represent single-character candidate words that connect every adjacent node as fallback candidates.The red, orange, and blue colors represent the components and lattice edges corresponding to the Neural, Statistical, and Rule-based segmenters, respectively.

3.2. Candidate Word Lattice Construction

Given an input sentence $S = c_1 c_2 \dots c_n$, we employ multiple base segmenters to generate initial segmentation results. Each segmenter outputs a sequence of non-overlapping words, which are uniformly mapped to candidate words covering specific character spans.

To ensure the completeness of the candidate space and to avoid fragmentation during decoding, we additionally include all single-character tokens in the sentence as fallback candidates. The final candidate word set for sentence S is defined as:

$$C(S) = \left(\bigcup_{k=1}^{K} U_k(S) \right) \cup U(S) \tag{1}$$

where $U_k(S)$ denotes the set of words generated by the k-th base segmenter for sentence S, and $U(S)$ denotes the set of all single-character substrings.

Based on the candidate set $C(S)$, we construct a directed acyclic word lattice. In this lattice, nodes correspond to character positions, and edges correspond to candidate words covering the associated character intervals. This structure provides a unified representation for subsequent candidate word scoring and global decoding.

3.3. Meta-Scoring Model

Once the word lattice is constructed, the segmentation task is reformulated as a confidence modeling problem over candidate words. For each candidate word $w_{i,j}$, the objective is to predict whether it should be included in the final segmentation result.

3.3.1. Feature Representation

Each candidate word is represented by a compact feature vector that jointly captures its source information, internal statistical properties, and local contextual structure.

First, source indicator features are introduced to denote whether a candidate word is generated by different base segmenters. Specifically, for each base segmenter, a binary feature indicates whether the candidate word is proposed by that segmenter. These features allow the meta-model to automatically learn the relative reliability of different segmenters under varying contextual conditions, obviating the need for heuristic rules or manual weight assignment.

Second, we incorporate unsupervised statistical features precomputed on large-scale unlabeled corpora to quantify the lexical plausibility of candidate words. Specifically, we adopt a pointwise mutual information (PMI)-inspired cohesion metric based on character

co-occurrence probabilities to measure the statistical association strength among characters within a candidate word. For a candidate word $w_{i,j} = c_i c_{i+1} \dots c_j$, its cohesion score is defined as:

$$\mathrm{PMI}(w_{i,j}) = \log \frac{P(w_{i,j})}{\prod_{t=i}^{j} P(c_t)} \tag{2}$$

where $P(w_{i,j})$ denotes the occurrence probability of $w_{i,j}$ in the unlabeled corpus, and $P(c_t)$ denotes the marginal probability of character c_t. This formulation serves as an approximation of standard pointwise mutual information and has been widely used in unsupervised new word discovery for Chinese.

In addition, to quantify the degree of boundary flexibility of candidate words, we introduce boundary entropy features based on adjacent character distributions. Specifically, we collect the sets of left-adjacent characters $N_L(w_{i,j})$ and right-adjacent characters $N_R(w_{i,j})$ of $w_{i,j}$ from the same unlabeled corpus, and define the left and right entropies as:

$$H_{\text{left}}(w_{i,j}) = -\sum_{x \in N_L(w_{i,j})} P(x \mid w_{i,j}) \log P(x \mid w_{i,j}) \tag{3}$$

$$H_{\text{right}}(w_{i,j}) = -\sum_{y \in N_R(w_{i,j})} P(y \mid w_{i,j}) \log P(y \mid w_{i,j}) \tag{4}$$

The overall boundary entropy of a candidate word span $w_{i,j}$ is defined as the sum of the two:

$$H(w_{i,j}) = H_{\text{left}}(w_{i,j}) + H_{\text{right}}(w_{i,j}) \tag{5}$$

All the above unsupervised statistical features are computed offline on independent large-scale corpora and do not rely on any gold-standard annotations during inference. As a result, they serve as external validation signals for the supervised model.

Beyond statistical features, we further incorporate a set of structural and context-related features to capture local boundary environments and positional cues of candidate words. These features include the character types of the immediate left and right neighbors, the relative position of the candidate word within the sentence, as well as the character types at the word boundaries. Together, they provide the meta-learner with lightweight yet informative morphological and contextual signals.

3.3.2. Meta-Learner

Based on the aforementioned feature representations, we adopt XGBoost [36], a scalable implementation of the Gradient Boosting Decision Tree (GBDT) algorithm, as the meta-learner to perform binary classification over candidate words. Specifically, it outputs the confidence probability $P(y = 1 \mid x_{i,j})$, where $x_{i,j}$ denotes the feature vector of candidate word $w_{i,j}$.

Unlike simple voting or static weighting strategies, the proposed meta-model dynamically adjusts its decision strategy according to contextual information, source agreement, and statistical properties of candidate words, enabling more fine-grained arbitration under diverse segmentation conflict scenarios.

3.4. Global Lattice Decoding

To ensure sentence-level consistency of the segmentation results, we perform global decoding over the word lattice. For each lattice edge corresponding to candidate word $w_{i,j}$, its weight is defined based on the confidence score produced by the meta-model:

$$\mathrm{score}(w_{i,j}) = \log\big(P(y = 1 \mid x_{i,j}) + \delta\big) \tag{6}$$

where δ is a smoothing constant introduced to ensure numerical stability. The logarithmic transformation converts the product of probabilities along a path into an additive form, enabling efficient global search via dynamic programming.

Let $dp(j)$ denote the maximum accumulated score from the beginning of the sentence to position j. The state transition equation is defined as:

$$dp(j) = \max_{(i,j)\in E} \left(dp(i) + \text{score}(w_{i,j}) \right) \tag{7}$$

where E denotes the set of edges in the word lattice. By backtracking the optimal path, the final segmentation covering the entire sentence can be obtained. This decoding procedure is formally equivalent to Viterbi search over the word lattice and effectively avoids boundary inconsistency issues caused by locally greedy decisions.

4. Experimental Results and Analysis

This section evaluates the SegFusion framework in alignment with the research objectives. We first establish a unified experimental protocol (Section 4.1), and then systematically analyze its performance across various scenarios through comparative experiments, ablation studies, and illustrative case analyses, with a particular focus on model complementarity, the role of unsupervised evidence in resolving low-consensus and OOV segmentation, and fine-grained behaviors reflected by word-length distributions (Sections 4.2–4.5).

4.1. Experimental Setup

Datasets and Evaluation Metrics. Experiments were conducted on four standard benchmark datasets from the SIGHAN Bakeoff 2005 (ICWB-2) for Chinese Word Segmentation (CWS), including PKU and MSR for Simplified Chinese, as well as CITYU and AS for Traditional Chinese. To eliminate discrepancies caused by different character sets, the Traditional Chinese corpora were converted into Simplified Chinese using nstools [37].

The computation of unsupervised statistical features, including Pointwise Mutual Information (PMI) and Branch Entropy, relies on large-scale external unlabeled corpora. In this work, a filtered Chinese Wikipedia corpus released on Hugging Face [38] was adopted as the statistical source to ensure data quality and reproducibility. Based on this corpus, character-level n-gram statistics were constructed and subsequently used for the calculation of unsupervised features.In practice, character n-grams up to length 5 were collected to balance statistical reliability and computational efficiency.

The evaluation metrics include Precision (P), Recall (R), F1-score (F_1), and Out-of-Vocabulary Recall (R_{OOV}). Their definitions are given as follows:

$$P = \frac{N_{\text{correct}}}{N_{\text{pred}}}, \quad R = \frac{N_{\text{correct}}}{N_{\text{gold}}}, \quad F_1 = \frac{2 \cdot P \cdot R}{P + R} \tag{8}$$

where N_{correct} denotes the number of correctly segmented words, N_{pred} denotes the total number of words predicted by the model, and N_{gold} denotes the total number of words in the gold standard annotations. For out-of-vocabulary (OOV) words, the recall is defined as:

$$R_{\text{OOV}} = \frac{N_{\text{OOV\_correct}}}{N_{\text{OOV\_total}}} \tag{9}$$

where $N_{\text{OOV\_total}}$ denotes the total number of OOV words in the test set that do not appear in the training set, and $N_{\text{OOV\_correct}}$ denotes the number of OOV words correctly segmented by the model.

Compared Methods and Baseline Settings. To comprehensively evaluate the effectiveness of SegFusion, representative methods at different levels were selected as comparison baselines, including the following categories.

(1) Heterogeneous Base Segmenters. Jieba was employed in precise mode with the HMM-based new word discovery enabled (HMM = True), using version 0.42.1. THULAC was executed with its default segmentation mode, using version 0.2.2. HanLP performed inference with the pretrained `FINE_ELECTRA_SMALL_ZH` model, using version 2.1.3.

(2) Static Ensemble Baseline. Majority Voting is a typical character-level hard voting ensemble method. It maps the outputs of different segmentation models into a unified character-level tagging scheme (e.g., the BMES scheme) and applies majority voting at the character level to determine the final segmentation results.

(3) Advanced Neural Baseline. WMSEG is a pretrained neural CWS model that enhances segmentation performance by incorporating multi-granularity n-gram wordhood information, and it is used as a performance reference for state-of-the-art neural approaches.

4.2. Main Experimental Results

To comprehensively evaluate the effectiveness of the proposed framework, the experiments were conducted under two settings: (1) *Cross-Paradigm Fusion*, which aims to verify the gains brought by heterogeneous integration; and (2) *SOTA Challenge*, which aims to compare the proposed method with one of the current state-of-the-art word segmentation systems.

4.2.1. Effectiveness of Heterogeneous Integration

As shown in Setting 1 of Table 1, SegFusion achieves stable and consistent performance improvements across all four datasets, validating the overall effectiveness of the heterogeneous integration framework.

In terms of overall performance, compared with the best-performing single model, SegFusion improves the F1-score by $+0.42\%$ to $+3.71\%$, with the most significant gain observed on the MSR dataset ($+3.71\%$). The MSR dataset contains a relatively high proportion of long compound words and named entities, where different segmentation paradigms are more likely to produce boundary disagreements. This result indicates that SegFusion can effectively integrate the complementary strengths of different models with respect to segmentation granularity. Moreover, compared with the majority voting baseline, SegFusion consistently outperforms it on all datasets, demonstrating that the feature-based dynamic arbitration mechanism exhibits stronger discriminative capability than simple voting strategies, particularly in low-consensus scenarios.

In terms of OOV recognition, SegFusion also achieves stable improvements, especially on the PKU and MSR datasets, where the OOV recall increases by $+3.78\%$ and $+4.10\%$, respectively. These results suggest that voting-based methods relying solely on model consensus suffer from inherent limitations in OOV scenarios, and that the introduction of unsupervised statistical features effectively alleviates this issue.

Table 1. Segmentation performance comparison under Cross-Paradigm Fusion and SOTA Challenge settings.

Method	PKU		MSR		AS		CITYU	
	F1	R_{OOV}	**F1**	R_{OOV}	**F1**	R_{OOV}	**F1**	R_{OOV}
Setting 1: Cross-Paradigm Fusion								
THULAC	92.28	79.15	85.41	43.76	84.60	62.08	85.16	71.38
HanLP	92.61	79.49	87.65	48.75	96.17	86.77	94.95	88.33
Jieba	81.83	58.26	81.18	44.86	82.16	53.27	82.38	68.36
BestSingle	92.61	79.49	87.82	48.75	96.17	86.77	94.87	88.33
MajorityVoting	94.57	81.48	88.30	47.40	92.36	73.61	91.98	80.95
SegFusion	**95.12**	**83.27**	**91.53**	**52.85**	**96.59**	**86.81**	**95.83**	**88.36**
Gain (Δ_{Best})	+2.51	+3.78	+3.71	+4.10	+0.42	+0.04	+0.88	+0.03
Setting 2: SOTA Challenge (+WMSEG)								
WMSEG	96.53	87.18	98.28	88.44	96.58	78.83	97.79	87.65
SegFusion	**96.61**	**87.43**	**98.35**	**88.54**	**96.59**	**78.89**	**97.81**	**87.69**
Gain (Δ_{SOTA})	+0.08	+0.25	+0.07	+0.10	+0.01	+0.06	+0.02	+0.04

4.2.2. Comparison with Advanced Systems

As shown in Setting 2 of Table 1, SegFusion achieves performance comparable to or slightly better than WMSEG across all four benchmark datasets. Specifically, on the PKU and MSR datasets, SegFusion attains F1-scores of 96.61% and 98.35%, respectively, both marginally higher than those of WMSEG. On the AS and CITYU datasets, the performance differences between the two methods remain small. These results indicate that even when an advanced neural segmentation system is included as a candidate model, complementary information among different model outputs still exists and can be further exploited.

It is worth emphasizing that SegFusion is a model-agnostic post-processing ensemble framework, which does not require additional annotated data or retraining of the underlying models, and thus offers strong generality while maintaining competitive performance.

4.3. Upper Bound and Complementarity Analysis

Table 2 summarizes the Oracle upper-bound results on the four datasets, which are used to assess the complementarity among base models and the theoretical improvement potential of ensemble methods. Specifically, Oracle Recall (R_{ora}) denotes the proportion of word boundaries that are correctly predicted by at least one base model; ΔR represents the improvement relative to the recall of the best single model; and S_{ora} denotes the proportion of sentences that can be perfectly segmented by combining all base model predictions.

Table 2. Oracle upper-bound results on four datasets.

Data	$F1_{best}$	R_{best}	R_{ora}	ΔR (pp)	S_{ora}
PKU	92.61	93.34	97.48	4.14	62.19
MSR	87.82	91.34	96.42	5.08	54.25
AS	96.17	97.25	99.06	1.81	93.24
CITYU	94.87	95.90	98.48	2.58	76.01

As shown in Table 2, the theoretical ensemble potential varies substantially across datasets. The MSR and PKU datasets exhibit larger upper-bound improvement margins. In particular, the Oracle Recall on MSR reaches 96.42%, leaving a margin of 5.08 percentage points over the best single model, while the corresponding improvement on PKU is

4.14 percentage points. These results indicate that significant prediction disagreements exist among base models on these datasets.

In contrast, the AS dataset shows a much smaller Oracle improvement of only 1.81 percentage points, and its Oracle sentence-level accuracy reaches 93.24%, suggesting that most samples are already correctly segmented by individual models or by model consensus, and that model complementarity is relatively limited. The CITYU dataset exhibits a moderate Oracle improvement margin of 2.58 percentage points. Notably, the Oracle sentence-level accuracy on MSR is only 54.25%, which is substantially lower than that of the other datasets, reflecting the presence of a large number of locally ambiguous and highly disputed samples. Overall, MSR and PKU theoretically offer greater ensemble potential, and subsequent analyses focus primarily on the MSR dataset.

4.4. Ablation Study

To analyze the contribution of each component in SegFusion, Table 3 reports the ablation results on the MSR dataset. It can be observed that removing any feature view leads to performance degradation, indicating that all components play a positive role in the ensemble decision process.

Table 3. Ablation results of different feature components in SegFusion on the MSR dataset.

Variant	F1	Δ	R_{OOV}	Δ
SegFusion (Full)	91.53	–	52.85	–
w/o Source	89.85	−1.68	49.35	−3.50
w/o Context	90.88	−0.65	51.45	−1.40
w/o Stats	91.01	−0.52	49.48	−3.37

From the perspective of overall F1-score, the source features contribute most significantly to performance improvements. Removing the source features results in a decrease of 1.68 percentage points in F1-score and 3.50 percentage points in ROOV, indicating that the predictions of different base models provide critical information for arbitration.

By comparison, context features have a relatively smaller impact on overall performance. When context features are removed, the F1-score decreases by only 0.65 percentage points and ROOV decreases by 1.40 percentage points, suggesting that these features mainly serve as auxiliary signals to enhance local decision stability.

Notably, unsupervised statistical features have a particularly strong impact on OOV recall. After removing the statistical features, the F1-score decreases by only 0.52 percentage points, while ROOV drops sharply by 3.37 percentage points. This result indicates that statistical features primarily contribute to more challenging OOV scenarios.

4.5. Fine-Grained Analysis

4.5.1. Low-Consensus Bucket Analysis

To further investigate the mechanism of SegFusion in OOV recognition, a stratified analysis of OOV samples was conducted on the MSR dataset based on model support. Support is defined as the number of base segmenters that correctly predict the gold-standard boundary of a given OOV word, and is used to characterize the predictability of OOV samples within the ensemble. Table 4 reports the OOV recall of different methods under different support levels, along with corresponding ablation results.

Table 4. Recall on OOV words stratified by support on the MSR dataset.

Support	OOV Count (Gold)	Voting	SegFusion	w/o Stats	Δ
1	293	4.44	62.80	44.50	−18.30
2	457	100.00	90.59	82.00	−8.59
3	895	100.00	100.00	100.00	0.00
Overall R_{OOV}	2829	–	52.85	49.48	−3.37

As shown in Table 4, the effect of unsupervised statistical features is primarily concentrated on low-support OOV samples with Support = 1. In this scenario, only one base model produces a correct prediction, while all others fail, causing majority voting to nearly collapse, with a recall of only 4.44%. In contrast, SegFusion significantly improves the OOV recall to 62.80%. When statistical features are removed, the recall drops to 44.50%, representing a decrease of 18.30 percentage points. This result demonstrates that PMI and branch entropy derived from external corpora provide critical complementary evidence for OOV recognition in highly disputed scenarios.

Under the Support = 2 condition, majority voting reaches its upper bound in OOV recall, while SegFusion performs slightly worse. This behavior indicates that SegFusion does not simply replicate voting decisions in this range, but instead makes more cautious path selections by jointly considering global structure and feature evidence. Unlike position-wise voting, SegFusion evaluates both local evidence and sentence-level consistency, thereby avoiding globally inconsistent segmentation paths. Under the Support = 3 condition, all methods achieve identical performance, indicating that no additional arbitration is required for such samples.

Overall, unsupervised statistical features significantly enhance SegFusion's discriminative capability on low-support OOV samples, while source and context features provide essential auxiliary support. By incorporating external statistical evidence, SegFusion is able to make more reliable arbitration decisions under high-disagreement conditions, effectively improving overall OOV recall.

4.5.2. Word-Length Bucket Analysis

To further analyze the behavior of SegFusion under different levels of structural complexity, we conduct a stratified analysis of segmentation results on the MSR dataset by word length and compare the recall performance of different methods on all words. Table 5 reports the results across different word-length intervals.

Table 5. Recall stratified by word length on MSR (All words).

Length	Gold	Voting	SegFusion	Δ
1	48,092	91.45	92.46	+1.01
2	49,472	94.73	96.29	+1.56
3	4652	77.67	85.40	+7.73
4	2711	52.67	66.77	+14.10
5+	1946	18.71	35.94	+17.23

From the word-length bucketed results in Table 5, it can be observed that SegFusion yields relatively limited improvements on short words, whereas its advantage increases substantially with word length. In particular, the recall gains become most pronounced for three-character words and longer, indicating that the main benefits of SegFusion do not stem from simple local boundary corrections, but rather from more effective modeling

of global structural consistency in long-word scenarios. By introducing a lattice-based representation, the segmentation process is transformed from position-wise local boundary decisions into a global path selection problem, thereby effectively alleviating the fragmented segmentation behavior that majority voting methods tend to exhibit in long-word and structurally complex cases.

When combined with the preceding analysis based on model support, it can be further observed that structurally complex long words are more likely to induce prediction disagreements among different segmentation models, and SegFusion demonstrates more pronounced advantages on such samples. This result explains, from the perspective of structural modeling, why SegFusion achieves stable performance gains in complex segmentation scenarios, and is consistent with the characteristics of low-consensus samples revealed by the support-based bucket analysis.

4.5.3. Illustrative Case Study

To further illustrate how SegFusion resolves segmentation disagreements in low-consensus and structurally complex scenarios, Table 6 presents a representative example comparing the segmentation results produced by different methods.

Table 6. An illustrative example of segmentation results produced by different methods.

Method	Segmentation
Jieba	东方红/三/号/卫星
THULAC	东方红/三号/卫星
HanLP	东方红三号/卫星
Voting	东方红/三号/卫星
SegFusion	东方红三号/卫星
Gold	东方红三号/卫星

This example corresponds to a low-consensus case (Support = 1), in which only one base segmenter (HanLP) predicts the correct boundary for the multi-character compound “东方红三号”. Due to insufficient agreement among base models, majority voting fails to preserve the complete word span and produces a fragmented segmentation. In contrast, SegFusion selects a globally consistent path at the lattice level and successfully recovers the correct segmentation, which is consistent with the structural complexity patterns discussed in the length-based analysis.

5. Discussion

In ensemble-based Chinese word segmentation, traditional static ensemble methods such as majority voting operate at the character level by making independent local boundary decisions and implicitly assume that prediction correctness correlates strongly with model consensus. This assumption often breaks down in ambiguous cases, where correct segmentations may only be supported by a small subset of base segmenters. SegFusion addresses this issue by reformulating ensemble learning as a word-level candidate re-scoring and global decoding problem. By jointly modeling heterogeneous model outputs, lightweight contextual cues, and the global structure of a word lattice, SegFusion enables sentence-level optimal path selection, thereby reducing error propagation in complex structures and yielding consistent improvements in overall F1 performance.

The Oracle upper-bound analysis further clarifies when such an ensemble framework is most beneficial. Datasets such as MSR and PKU exhibit larger Oracle margins, indicating substantial disagreement and stronger complementarity among base segmenters, and thus provide greater room for ensemble optimization. In contrast, when sentence-level accuracy

is already high and predictions from different models are largely consistent, the achievable gains of ensemble methods are naturally constrained by the collective coverage of the base models. This observation highlights that the effectiveness of SegFusion is closely tied to the diversity and structural complementarity of the selected heterogeneous segmenters.

Within this framework, unsupervised statistical features, including mutual information and branching entropy, play a targeted corrective role, particularly in low-consensus scenarios. Support-based analysis shows that when only one base segmenter predicts the correct boundary (Support = 1), majority voting nearly collapses, whereas SegFusion can recover many of these difficult out-of-vocabulary (OOV) cases by leveraging model-independent statistical evidence derived from large-scale unlabeled corpora. This behavior is also consistent with the ablation results, in which removing statistical features leads to a pronounced degradation in R_{OOV}, confirming their critical contribution under high-disagreement conditions.

Moreover, the word-length analysis provides complementary structural evidence for the effectiveness of SegFusion. As word length increases, local boundary disagreements tend to accumulate, making character-level voting methods increasingly prone to over-segmentation and fragmentation of long compound words. By evaluating candidate words as coherent units within the lattice and selecting globally consistent paths, SegFusion preserves the structural integrity of multi-character words. Consequently, its advantages become more pronounced on datasets dominated by long words and complex entity structures, such as MSR.

Nevertheless, this study has several limitations. First, the effectiveness of the unsupervised statistical features depends on the scale and coverage of the external unlabeled corpora, as these features rely on reliable frequency statistics to distinguish plausible candidate words. Second, the overall performance upper bound of SegFusion is constrained by the candidate word space generated by the participating base segmenters. When different segmenters exhibit greater diversity in segmentation granularity and structural preferences, the resulting candidate pool becomes richer and the lattice offers more flexibility for global optimization; conversely, when base models produce highly similar segmentations, the search space of the ensemble is inherently limited, restricting further performance improvements.

6. Conclusions and Future Work

This paper addresses the widespread disagreement among different Chinese word segmentation models in scenarios involving complex ambiguities and out-of-vocabulary words, and proposes SegFusion, a lattice-based dynamic meta-arbitration ensemble framework that integrates unsupervised statistical features. The proposed method performs unified re-scoring over word-level candidates generated by multiple heterogeneous segmenters and derives consistent segmentation results through global lattice decoding, thereby alleviating the performance degradation of traditional character-level voting methods in low-consensus scenarios.

Experimental results on multiple standard benchmark datasets demonstrate that SegFusion consistently outperforms individual models and static ensemble baselines in terms of both overall segmentation performance and OOV recall. In particular, on the highly ambiguous MSR dataset, SegFusion achieves improvements of 3.71% in F1 score and 4.10% in OOV recall. Further fine-grained analyses show that unsupervised statistical features provide critical model-independent evidence for resolving low-consensus OOV cases (Support = 1), while lattice-based global structural modeling effectively mitigates the instability of local boundary decisions in long-word and structurally complex scenarios. These results validate the effectiveness of combining word-level global structural modeling

with external linguistic statistical information in the ensemble arbitration stage for Chinese word segmentation.

Future work will focus on two directions. On the one hand, we plan to incorporate a more diverse set of base segmenters to further enrich the candidate lattice space and raise the ensemble upper bound. On the other hand, we aim to extend the proposed framework to cross-domain segmentation scenarios, in order to investigate the robustness of unsupervised statistical features under limited labeled data conditions.

Author Contributions: Conceptualization, C.W. and J.D.; Methodology, C.W.; Software, C.W.; Formal analysis, C.W.; Investigation, C.W.; Writing—original draft preparation, C.W.; Writing—review and editing, C.W. and J.D.; Supervision, J.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Postgraduate Research Innovative Project of Central South University (Grant No. 2025ZZCX0558).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available and commonly used benchmark corpora in Chinese word segmentation research.

Acknowledgments: During the preparation of this manuscript, the author used ChatGPT (version 5.2, OpenAI) for language polishing and academic writing refinement. The author has reviewed and edited the content and takes full responsibility for the publication.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

CWS	Chinese Word Segmentation
OOV	Out-of-Vocabulary
PMI	Pointwise Mutual Information
MSR	Microsoft Research Corpus
PKU	Peking University Corpus
AS	Academia Sinica Corpus
CITYU	City University of Hong Kong Corpus

References

1. Mai, C.C.; Chen, Y.; Gong, Z.; Wang, H.; Qiu, M.; Yuan, C.; Huang, Y. PromptCNER: A Segmentation-Based Method for Few-Shot Chinese NER with Prompt-Tuning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2025**, *24*, 1–24. [CrossRef]
2. Ye, C.; Hernandez, A.A.; Abisado, M. Chinese Named Entity Recognition Based on Lexicon Knowledge Enhancement. *J. Inf. Hiding Multimed. Signal Process.* **2025**, *16*, 917–931.
3. Lv, S.; Ding, X. Chinese Relation Extraction with External Knowledge-Enhanced Semantic Understanding. *Int. J. Adv. Comput. Sci. Appl.* **2025**, *16*, 1317. [CrossRef]
4. Chen, Y.; Li, Z.; Zhang, C.; Yang, C.; Cady, A.; Lee, A.K.; Zeng, Z.; Jo, E.L.; Pan, H.; Park, J. Parsing through Boundaries in Chinese Word Segmentation. *arXiv* **2025**, arXiv:2503.23091. [CrossRef]
5. Hu, C.; Si, Q.; Wang, S. Research on Dynamic Curriculum Learning in Mongolian-Chinese Neural Machine Translation. In Proceedings of the 2025 International Joint Conference on Neural Networks (IJCNN), Rome, Italy, 30 June–5 July 2025; pp. 1–8.
6. Zhang, Y.; Yang, J. Chinese NER Using Lattice LSTM. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 1554–1564.
7. Huang, C.R.; Šimon, P.; Hsieh, S.K.; Prévot, L. Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Wordbreak Identification. In Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, Czech Republic, 25–27 June 2007; pp. 69–72.
8. Du, G. Research Advances in Chinese Word Segmentation Methods and Challenges. *Appl. Comput. Eng.* **2024**, *37*, 16–22. [CrossRef]

9. Sun, W.; Wan, X. Reducing Approximation and Estimation Errors for Chinese Lexical Processing with Heterogeneous Annotations. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jeju, Republic of Korea, 8–14 July 2012; pp. 232–241.
10. Min, K.; Ma, C.; Zhao, T.; Li, H. BosonNLP: An Ensemble Approach for Word Segmentation and POS Tagging. In *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC)*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 520–526.
11. Fu, J.; Liu, P.; Zhang, Q.; Huang, X.J. RethinkCWS: Is Chinese word segmentation a solved task? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 5676–5686.
12. Chen, K.J.; Liu, S.H. Word Identification for Mandarin Chinese Sentences. In Proceedings of the 14th International Conference on Computational Linguistics (COLING 1992), Nantes, France, 23–28 August 1992.
13. Eddy, S.R. Hidden Markov Models. *Curr. Opin. Struct. Biol.* **1996**, *6*, 361–365. [CrossRef] [PubMed]
14. McCallum, A.; Freitag, D.; Pereira, F.C.N. Maximum Entropy Markov Models for Information Extraction and Segmentation. In Proceedings of the 17th International Conference on Machine Learning (ICML), Stanford, CA, USA, 29 June–2 July 2000; pp. 591–598.
15. Lafferty, J.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the 18th International Conference on Machine Learning (ICML), Williamstown, MA, USA, 28 June–1 July 2001; pp. 282–289.
16. Yao, Y.; Huang, Z. Bi-directional LSTM Recurrent Neural Network for Chinese Word Segmentation. In *Proceedings of the International Conference on Neural Information Processing (ICONIP)*; Springer: Berlin/Heidelberg, Germany, 2016, pp. 345–353.
17. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
18. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z. Pre-training with Whole Word Masking for Chinese BERT. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3504–3514. [CrossRef]
19. Diao, S.; Bai, J.; Song, Y.; Zhang, T.; Wang, Y. ZEN: Pre-training Chinese Text Encoder Enhanced by N-gram Representations. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020; pp. 4729–4740.
20. Liu, W.; Fu, X.; Zhang, Y.; Xiao, W. Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), Dublin, Ireland, 22–27 May 2022.
21. Su, Y.; Liu, F.; Meng, Z.; Lan, T.; Shu, L.; Shareghi, E.; Collier, N. TaCL: Improving BERT Pre-training with Token-Aware Contrastive Learning. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, USA, 10–15 June 2022; pp. 2497–2507.
22. Zhang, Z.; He, L.; Li, Z.; Zhang, L.; Zhao, H.; Du, B. Segment First or Comprehend First? Explore the Limit of Unsupervised Word Segmentation with Large Language Models. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, Vienna, Austria, 27 July–1 August 2025; pp. 18146–18163.
23. Chu, D.; Tan, Z.; Wan, B.; Fang, F.; Zhou, S.; Zhu, Y.; Zhu, M.; Wu, Y. LLM-GeoCWS: A Semi-Supervised Chinese Word Segmentation Method Using Large Language Model for Geoscience Domain. *Earth Sci. Inform.* **2025**, *18*, 525. [CrossRef]
24. Tang, M.T.; Mi, C.G. Improving Ancient Chinese Word Segmentation with Knowledge-Enhanced Prompting for Large Language Models. *Int. J. Intell. Syst.* **2025**, *2025*, 9612240. [CrossRef]
25. Yang, J.; Zhang, Y.; Liang, S. Subword Encoding in Lattice LSTM for Chinese Word Segmentation. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, MN, USA, 2–7 June 2019; pp. 2720–2725.
26. Wu, S.; Song, X.; Feng, Z.; Wu, X.J. NFlat: Non-flat Lattice Transformer for Chinese Named Entity Recognition. *arXiv* **2022**, arXiv:2205.05832.
27. Qiu, X.; Pei, H.; Yan, H.; Huang, X.J. A Concise Model for Multi-Criteria Chinese Word Segmentation with Transformer Encoder. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16–20 November 2020; pp. 2887–2897.
28. Tian, Y.; Song, Y.; Xia, F.; Zhang, T.; Wang, Y. Improving Chinese Word Segmentation with Wordhood Memory Networks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, 5–10 July 2020; pp. 8274–8285.
29. Xu, Z.; Xiang, Y. Word Structure Embedding and In-Context Learning for Chinese Segmentation. *IEEE Access* **2025**, *13*, 207617–207637. [CrossRef]
30. Song, D.; Sarkar, A. Voting between Dictionary-Based and Subword Tagging Models for Chinese Word Segmentation. In Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, 22–23 July 2006; pp. 126–129.

31. Zheng, Z.; Wang, Y.; Lepage, Y. Chinese Word Segmentation Based on Analogy and Majority Voting. In Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC): Posters, Shanghai, China, 30 October–1 November 2015; pp. 151–156.
32. Sun, W. Word-Based and Character-Based Word Segmentation Models: Comparison and Combination. In Proceedings of the COLING 2010: Posters, Beijing, China, 23–27 August 2010; pp. 1211–1219.
33. Gao, Q.; Vogel, S. A Multi-layer Chinese Word Segmentation System Optimized for Out-of-domain Tasks. In Proceedings of the CIPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing, China, 28–29 August 2010.
34. Sun, W. A Stacked Sub-word Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), Portland, OR, USA, 19–24 June 2011; pp. 1385–1394.
35. Xu, J.; Sun, X.; Li, S.; Cai, X.; Wei, B. Deep Stacking Networks for Low-Resource Chinese Word Segmentation with Transfer Learning. *arXiv* **2017**, arXiv:1711.01427. [CrossRef]
36. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
37. Chen, S. nstools: A Toolkit for Chinese Text Normalization. GitHub Repository. 2010. Available online: https://github.com/skydark/nstools (accessed on 5 January 2026).
38. wikipedia-cn-20230720-Filtered. Hugging Face Dataset. 2023. Available online: https://huggingface.co/datasets/pleisto/wikipedia-cn-20230720-filtered (accessed on 5 January 2026).