Span-based Multi-grained Word Segmentation with Natural Annotations

Anonymous ACL submission

Abstract

Multi-grained word segmentation (MWS) differs from traditional single-grained word segmentation (SWS) by dividing a sentence into multiple word sequences at varying granularities. The scarcity of annotated MWS data has led previous studies to use automatically generated pseudo MWS data and treat MWS as a tree parsing task. However, this method is limited by the low quality of the pseudo data. In this work, we directly utilize multiple single-grained datasets and implement multi-task learning for MWS. To better address conflicts arising from words segmented at different granularities, we employ a span-based word segmentation model. Additionally, we incorporate naturally annotated BAIKE data to improve model performance in cross-domain applications. Experimental results demonstrate that our method achieved an F1 score improvement of 0.83 on the NEWS dataset and 4.8 on the BAIKE dataset. Furthermore, by employing data augmentation, we obtained an additional F1 score improvement of 2.23 on the BAIKE dataset.

1 Introduction

Chinese word segmentation (CWS) plays a crucial role in natural language processing (NLP). Over the past decade, CWS has made significant progress (Zhao et al., 2017, 2018b; Shi et al., 2019; Yang, 2019; Li et al., 2023; Xu, 2024). Unlike English, Chinese lacks clear word boundaries, and different individuals have varying perceptions of word boundaries. Therefore, single-grained word segmentation (SWS) cannot fully meet the segmentation needs of Chinese. Multi-grained word segmentation (MWS) enriches semantic information by acquiring both coarse-grained and fine-grained word boundary details, enhancing its adaptation to diverse NLP tasks. For instance, Dou et al. (2024) successfully integrated MWS and named entity recognition (NER), resulting in notable improvement in entity identification.

现任龙岩市国土资源局党组成员、副局长					
CTB	龙岩	市	国土	资源	局
MSR	龙岩市		国土资源局		
PKU	龙岩市		国土	资源局	
	Longyan	City	Nationl Land	Resources	Bureau

Table 1: The different segmentation results of the natural annotation segment under three annotation specifications.

The current research on MWS follows two main approaches, both of which believe that sentences can be segmented at different granularities. One approach, which refers to itself as MCCWS (Multi-Criteria Chinese Word Segmentation) (Chen et al., 2017; Gong et al., 2019; Qiu et al., 2020), learns multiple segmentation criteria during training but selects the most appropriate criterion as the output during prediction, resulting in a final segmentation that is still of a single granularity.

The other approach, proposed by Gong et al. (2017), treats MWS as a structured prediction task where all levels of granularity are retained simultaneously during the prediction process. To overcome the challenge of limited training data, coupled models are employed to merge two segmentation annotations into combined annotations, ultimately resulting in constituent trees annotated across all granularities. However, the conversion process is intricate, raising doubts about the quality of the generated pseudo-data. Moreover, the tree parsing necessitates the CKY algorithm (Kasami, 1965; Younger, 1967) during inference, leading to a significant increase in time complexity. Consequently, this approach encounters difficulties in efficiently achieving the objectives of MWS.

In this work, we build upon the work of Gong et al. (2017), addressing the challenge of insufficient standard training data in MWS by enabling the joint learning of multiple segmentation granularities. Our approach leverages manually annotated SWS data to train the model and produces MWS results with a span-based word segmentation (WS) model. We selected three classic datasets from the field of Chinese word segmentation: the Penn Chinese Treebank (CTB) (Xue et al., 2005), the Microsoft Research Chinese Word Segmentation (MSR) corpus (Huang et al., 2006), and the People's Daily Corpus (PKU) from Peking University (Yu and Zhu, 1998). The CTB prefers fine-grained annotation, making it more suitable for syntactic and semantic analysis. The MSR corpus provides coarse-grained annotation, typically identifying named entities as complete words. The PKU corpus lies between the two, with coarse-grained annotation aiding in information retrieval and extraction tasks. While our method can efficiently utilize these data, we noticed that segmentations at different granularities could result in conflicts during the decoding process, with 1.7% of segmentations in the test set showing such problems. To mitigate this, we introduced a CKY decoding module specifically designed to resolve these conflicts. Additionally, we incorporated naturally annotated BAIKE data and used marginal probabilities to select high-quality training examples, thereby enhancing the model's performance on cross-domain test data.

Our main contributions can be summed as follows:

1. We use a span-based segmentation model to leverage SWS data for MWS. The semi-Markov algorithm is employed for efficient training and prediction.
2. We introduce a CKY decoding module to address conflicts in MWS and select the optimal segmentation results. The conflict resolution leads to improvements of 0.63 and 1.68 F-scores on the NEWS-test and BAIKE-test respectively.
3. We introduce naturally annotated BAIKE data for cross-domain MWS. Through learning from partially labeled natural texts, the model achieves a maximum F1 improvement of 2.23 on the BAIKE-test.

2 Related Work

MWS approaches. Gong et al. (2017) first proposal the concept of MWS and used automatically generated pseudo-data to train a tree parser. Subsequently, researchers have used SWS data and dictionary data as additional weak label training data to further enhance MWS performance (Gong et al., 2020). Additionally, some scholars are dedicated to the research of MCCWS. They believe that Chinese text segmentation involves multiple criteria, with each sentence having an optimal criterion. To address this, they have sequentially employed Multi-Task Learning (MTL) (Chen et al., 2017; Gong et al., 2019) and Unified Model approaches (Qiu et al., 2020), aiming to identify the most suitable criterion through input cues. Chou et al. (2023) proposed using adversarial multi-criteria learning to leverage the shared knowledge across multiple heterogeneous criteria to improve performance under a single criterion.

Utilizing weakly labeled data. Jiang et al. (2013) trained the enhanced classifier on weakly labeled web data by using the annotation differences between the outputs of constraint decoding and normal decoding. Liu et al. (2014) and Zhao et al. (2018a) utilized various sources of free annotated data, combining fully and partially annotated data to train the model, demonstrating the effectiveness of free data.

Gong et al. (2020) using naturally annotated data from dictionaries for the MWS task. However, the concise specifications of dictionary data led to minimal gains obtained by the model.In this paper, we using naturally annotated segments from the Baidu Baike data, we obtain their MWS results for model training.

Span-based methods. In the early stages of CWS, reliance was primarily on manually curated dictionaries and rules (Zhao et al., 2018b). As computational capabilities advanced, statistical methods such as Conditional Random Fields (CRF) began to be introduced (Peng et al., 2004; Sutton et al., 2007; Liu et al., 2016; Jin et al., 2022). In recent years, the emergence of various pre-trained models has enabled the capture of richer contextual information, achieving excellent results in word segmentation (Li et al., 2022). Span-based methods (Wang et al., 2022) can be regarded as a variant of sequence labeling methods (Shin and Lee, 2020; Xue, 2003), enabling more direct modeling and prediction of word boundaries in text.

Algorithm 1 Semi-Markov Algorithm.

Input: Sentence $\boldsymbol{x} = c_1 c_2 \ldots c_n$ and span scores $s(i, j)$ for each candidate word $\boldsymbol{x}_{i:j}$
Define: $\alpha \in \mathbf{R}^{n+1}$ stores the highest score of the partial segmentation results
Initialize: $\alpha[0] = 0$
for $j = 1 \ldots n$ **do**
  $\alpha[j] = \max_{\max(1, j-M) \leq i \leq j} \alpha[i-1] + s(i, j)$
  ▷ *M is maximum word length* ◁
return $\alpha[n]$

3 Span-based Word Segmentation

3.1 Task Definition

Formally, a CWS model divides a character sequence $\boldsymbol{x} = c_1 c_2 \ldots c_n$ into a word sequence $\boldsymbol{y} = w_1 w_2 \ldots w_m$, where $w_k = \boldsymbol{x}_{i:j}$ is the kth word spanning from character c_i to character c_j.

In this work, we utilize a span-based model built on semi-Markov conditional random fields (semi-CRFs) for CWS. Each word $w_k = \boldsymbol{x}_{i:j}$ is assigned a score $s(i, j)$, and the segmentation score of $\boldsymbol{y}$ is the sum of scores of all words:

$$s(\boldsymbol{x}, \boldsymbol{y}) = \sum_{w_k = \boldsymbol{x}_{i:j} \in \boldsymbol{y}} s(i, j) \tag{1}$$

Under the semi-CRF framework, the conditional probability of the segmentation result $\boldsymbol{y}$ given the input $\boldsymbol{x}$ is defined as:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp(s(\boldsymbol{x}, \boldsymbol{y}))}{Z(\boldsymbol{x}) \equiv \sum_{\boldsymbol{y}' \in \mathcal{Y}} \exp(s(\boldsymbol{x}, \boldsymbol{y}'))} \tag{2}$$

where $Z(\boldsymbol{x})$ is the normalization term and $\mathcal{Y}$ represents the set of all possible segmentation results for $\boldsymbol{x}$.

3.2 Inference Algorithm

Given the scores of all candidate words, the goal is to find the optimal segmentation result $\hat{\boldsymbol{y}}$, which achieves the highest segmentation score:

$$\hat{\boldsymbol{y}} = \operatorname*{argmax}_{\boldsymbol{y} \in \mathcal{Y}} s(\boldsymbol{x}, \boldsymbol{y}) \tag{3}$$

The inference process can be efficiently addressed using the semi-Markov algorithm. This algorithm processes the input $\boldsymbol{x}$ sequentially from left to right to derive a partially optimal segmentation. Please refer to Algorithm 1 for details. The computational complexity of the semi-Markov algorithm is initially $O(n^2)$, but can be reduced to $O(Mn) = O(n)$ by constraining the maximum word length to M.

3.3 Training Loss

The loss function is defined as the negative log-likelihood (refer to Equation 2) of gold-standard segmentation $\boldsymbol{y}^*$:

$$\mathcal{L}(\boldsymbol{x}) = -\log p(\boldsymbol{y}^*|\boldsymbol{x}) = \log Z(\boldsymbol{x}) - s(\boldsymbol{x}, \boldsymbol{y}^*) \tag{4}$$

where $Z(\boldsymbol{x})$ is calculated using the semi-Markov algorithm by replacing the max-product with sum-product. The marginal probability of each word is derived through the partial derivative of $\log Z(\boldsymbol{x})$ with respect to the word score $s(i, j)$. This marginal probability is subsequently utilized in conflict resolution (see Subsection 4.3) and data selection (refer to Subsection 5.1).

4 Multi-grained Word Segmentation

Let $\boldsymbol{x} = c_1 c_2 \ldots c_n$ be a character sequence of length n. The goal of MWS is to produce multiple segmentation results for $\boldsymbol{x}$, corresponding to different segmentation standards or granularities. For each granularity g, the segmentation result is $\boldsymbol{y}^g = w_1^g w_2^g \ldots w_m^g$. In this work, instead of choosing a specific granularity g^* as done in other research (Chou et al., 2023), we adopt the approach proposed by Gong et al. (2017), where all granularities are retained, creating a hierarchical structure (refer to Table 1). Although this method is generally effective, there are instances where words from different granularities overlap, hindering the establishment of a coherent hierarchical structure. In the following section, we will address:

1. For a given input $\boldsymbol{x}$, how to obtain segmentation results of three annotation granularities (in this work $g \in$ {CTB, PKU, MSR}).
2. How to resolve conflicts arising from segmentation under different standards.

4.1 Span-based MWS

MWS can provide more richer information than SWS. Despite its benefits, the primary obstacle is the scarcity of training data due to the absence of an established multi-grained word segmentation dataset. Gong et al. (2017) developed a synthetic dataset for multi-grained word segmentation by combing multiple segmentation annotations of sentences into constituent trees. Nonetheless, this process necessitates a complicated conversion procedure which may introduce inaccurate examples. Additionally, the inference stage relies on the CKY algorithm, which operates at a time complexity of $O(n^3)$.

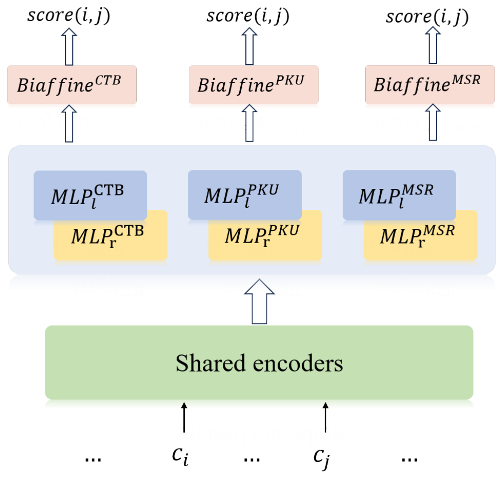

Figure 1: Model architecture.

To address these challenges, this work employs a multi-task learning (MTL) strategy for MWS by treating each segmentation granularity as an individual task. Figure 1 illustrates the setup, where the CTB, PKU, and MSR datasets share one encoder while utilizing three distinct decoders. For each sentence belonging to a specific granularity, the contextual representations from the shared encoder are input to the granularity-specific decoder to compute the loss during the training phase. In the inference stage, the three decoders independently predict segmentation results, which are subsequently arranged into a hierarchical structure. The training and inference procedures leverage the span-based word segmentation model detailed in Section 3. Compared to Gong et al. (2017), our method provides two main advantages:

1. The model is trained on sentences with word segmentation labels at a single granularity, eliminating the need for multi-granularity annotations.
2. While each decoder needs to conduct word segmentation to produce multi-granularity results, incurring a computational cost of $O(3n^2)$, it remains more efficient than the CKY-based method.

4.2 Model Framework

In this work, we constructed a span-based MTL model as shown in Figure 1, where the encoder is shared across multiple granularities while maintaining separate decoders for each granularity. The whole network architecture is similar to Zhang et al. (2020).

Encoder. For each input character c_i, a shared encoding layer is used to encode it and obtain the contextual representation h_i.

Boundary representation. Within each decoder, two MLP layers are used to obtain the left and right boundary representation vectors for each character c_i.

$$\begin{aligned} h_i^l &= \mathbf{MLP}^l(h_i) \\ h_j^r &= \mathbf{MLP}^r(h_j) \end{aligned} \tag{5}$$

Biaffine scoring. The representations of the left and right boundaries are then passed through a Biaffine layer to compute the score $s(i, j)$ for each word.

$$s(i,j) = \begin{bmatrix} h_i^l \\ 1 \end{bmatrix}^T \mathbf{W}(h_j^r) \tag{6}$$

Decoding. Subsequently, the scores of candidate words are input to the semi-Markov algorithm to derive granularity-specific segmentation results.

4.3 Conflict Resolution

Since MTL framework separately predicted segmetations of each granularity, conflicts may arise when words from different granularities overlap. A formal definition of *conflict* is provided herein For any two words $\boldsymbol{x}_{i:j}$ and $\boldsymbol{x}_{s:t}$, we say they overlap if $s \leq j < t$ when $i < s$ or $s < i \leq t$ when $j > t$. In the SWS and MCCWS approachs, with only one segmentation sequence, conflicts are avoided. However, in the MWS method, potential conflicts hinder the formation of legally hierarchical segmentation outputs. There are two common methods for resolving conflicts in the existing literature:

1. **Ignoring Conflicts**: Given that the number of conflicts is very small (1,7%[1] in this work), conflicts are not addressed, and all overlapping words are retained in the final outputs (Gong et al., 2017).
2. **Voting Mechanism**: Different segmentation granularities are voted on to select words with the highest votes (Saha and Ekbal, 2013). However, when two words receive the same number of votes, this method randomly chooses one.

This work introduces an optimal conflict resolution strategy using a span-based model, which is challenging for sequence labeling models to achieve the same objective. Given that hierarchical

[1]We count the number of conflicts in the test set and calculate the proportion of these conflicts relative to the total word counts.

structures are fundamentally trees, the span-based model offers all the necessary components, such as span scores, to leverage the CKY algorithm for identifying the highest-scoring tree structures. Initially, we calculate the marginal probability of each span under three granularities using marginal inference (refer to Subsection 3.3). Subsequently, we select the maximum probability[2] from three granularities for each span as input for the CKY algorithm. For any two conflicting words, we only reserve the one appears in the output trees.

5 Data Augmentation with Natural Annotations

In the study by Gong et al. (2020), it was observed that the MWS model shows good performance on newswire data but experiences a notable decrease in accuracy when applied to the cross-domain BAIKE data (akin to Wikipedia). This drop in performance is attributed to the substantial differences between the training and testing data. To tackle this issue, they sought to improve the model by incorporating weakly labeled data from dictionary resources to gain insights into word boundary information. However, the dictionary resources presented two clear limitations: 1) the words were predominantly short, mainly consisting of two-character words; 2) they did not align with the domain of BAIKE. To overcome these shortcomings, this paper suggests utilizing BAIKE data for data augmentation. BAIKE data comes with naturally annotated spans like anchor texts that frequently denote entities and phrases, thereby offering rich granularity information.

5.1 Data Filtering

While naturally annotated spans can provide valuable information, noise may be introduced due to its differences with source-domain segmentation standards (Liu et al., 2014). To deal with this issue, we choose utilizing the probability of these spans to select high-quality training examples. Specifically, we first categorize sentences into distinct probability intervals based on the probability of naturally annotated spans.[3] Although lower probability intervals generally indicate lower quality, we employ two metrics to assess and determine the appropriate interval:

[2]We also tried using the averaged probability, but it yielded inferior results.

[3]The probability is also calculated with marginal inference described in Subsention 3.3

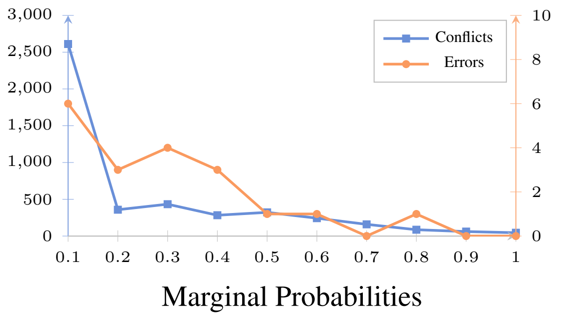

Figure 2: The criteria for data filtering, with the left side indicating the number of conflicts and the right side indicating the number of prediction errors.

1. We quantify the conflicts between the predicted outcomes of the MWS model and the naturally annotated spans within each probability interval.
2. We randomly sample 100 sentences from each probability interval and assess the number of mispredicted sentences through human evaluation.

As illustrated in Figure 2, intervals with probabilities exceeding 0.4 exhibit a notable decrease in both the number of automatically evaluated conflicts and manual inspected errors. Consequently, these data will be utilized as high-quality training samples. Our experiments will further investigate the influence of varying data scales and probability intervals on model performance.

5.2 Obtain Partial Multi-grained Annotation

For span-based WS model, natural annotations cannot be directly used for training because words serve as fundamental elements in semi-Markov algorithm. Thus, it is essential to obtain segmentation annotations at different granularities.

Similar to the self-training method, we employ the MWS model to predict multi-grained results, which are further used as gold-standard annotations. Notably, only the segmentation results corresponding to the naturally annotated spans are retained. For example, as shown in Table 1 , the original sentence "现任龙岩市国土资源局党组成员、副局长" contains the natural annotation "龙岩市国土资源局", which is segmented into three granularities. We solely preserve the segmentation outcomes for this annotated text while disregarding the segmentation of other sentence components.

Furthermore, to mitigate potential inaccuracies stemming from erroneous predictions, we discard sentences where conflicts occur. These conflicts

Probability	#Sent	#Spans	Conflict(%)	CTB		MSR		PKU	
				#spans	inc(%)	#spans	inc(%)	#spans	inc(%)
0.4-0.5	350k	550K	0.16	675K	21.68	700K	27.23	661K	20.13
0.5-0.6	450K	679K	0.11	791K	16.61	810K	19.35	768K	13.31
0.6-0.7	700K	1.01M	0.06	1.12M	10.89	1.00M	-	1.08M	6.93
0.7-0.8	650K	974K	0.05	1.04M	6.75	1.03M	5.73	1,02M	4.70
0.8-0.9	650K	1.11M	0.03	1.15M	3.60	1.14M	2.70	1.14M	2.70
0.9-1.0	2.4M	3.27M	0.02	3.37M	3.06	3.29M	0.61	3.35M	2.45

Table 2: Data statistics used in the process of data augmentation (K represents thousand and M represents million). We calculated the number of sentences (#Sent) in each probability interval, the number of naturally annotated segments (#Spans), the proportion of conflicts (Conflicts), and the number of segmentations (#span) obtained under the three annotation standards. (inc) indicates the proportion of span count increase under the corresponding annotation standard.

encompass instances where predicted words clash with naturally annotated spans and where words segmented at different granularities contradict each other.

5.3 Training with Partial Annotation

The obtained BAIKE sentences are mixed with source-domain sentences to enhance the MWS moded. In the training phase, these BAIKE sentence are fed into three decoders (see Figure 1) for each granularity to compute three losses, which are then summed up as the final loss. The primary challenge lies in calculating the loss when only partial annotations are available, prompting us to employ the CRF model as a solution. We will use one granularity as an example to demonstrate how this can be accomplished.

Let $\widetilde{\boldsymbol{x}} = c_i...c_j$ represent a naturally annotated span, which is a part of a sentence $\boldsymbol{x} = c_1 \dots \widetilde{\boldsymbol{x}} ... c_n$. The sentence only contains partial annotation information $\widetilde{\boldsymbol{y}} = w_s \dots w_t$ corresponding to $\widetilde{\boldsymbol{x}}$. We say $\boldsymbol{y}^* = w_1 ... \widetilde{\boldsymbol{y}} ... w_m$ is a complete segmentation of $\widetilde{\boldsymbol{y}}$, and the state space $\widetilde{\mathcal{Y}}$ consists of all such segmentations. The normalized probability of $\widetilde{\mathcal{Y}}$ is defined as:

$$p(\widetilde{\boldsymbol{y}}|\boldsymbol{x}) = \frac{\widetilde{Z}(\boldsymbol{x}) \equiv \sum_{\boldsymbol{y}^* \in \widetilde{\mathcal{Y}}} \exp(s(\boldsymbol{x}, \boldsymbol{y}))}{Z(\boldsymbol{x}) \equiv \sum_{\boldsymbol{y}' \in \mathcal{Y}} \exp(s(\boldsymbol{x}, \boldsymbol{y}'))} \quad (7)$$

The training objective of the model is to find as many complete segmentation as possible for partial annotation $\widetilde{\boldsymbol{y}}$ and maximize this probability. The calculation of the loss function is as follows:

$$\mathcal{L}(\boldsymbol{x}) = -\log p(\widetilde{\boldsymbol{y}}|\boldsymbol{x}) = \log Z(\boldsymbol{x}) - \log \widetilde{Z}(\boldsymbol{x}) \quad (8)$$

6 Experiments

Data. We use three datasets across different granularities: CTB, MSR, and PKU. For evaluation, we employ NEWS-dev, NEWS-test and BAIKE-test provided by Gong et al. (2020). Additionally, for cross-domain data augmentation, we filtered and acquired 5.2 million naturally annotated examples from 12 million sentences. Data statistics are shown in Table 2 and Table 3.

Settings. Following Gong et al. (2017), we use standard measures of F1, precision (P), and recall (R) scores to evaluate MWS. Two types of encoder are used: BiLSTM and BERT[4] (Devlin et al., 2019). The configuration of the model adheres to Zhang et al. (2020). The training epochs are set to 1000 and 15 respectively, and early stopping is applied on the development set.

6.1 Benchmark Methods

We employ five methods for comparison. Alongside the multi-task learning method proposed in this work, we replicated two benchmark methods: tree parsing and single-task learning.

1. **Tree-based**: Gong et al. (2017) use pseudo MWS data in the form of constituent trees to train a tree parser with CKY decoding. We reproduce their method when BERT is used.
2. **Single**: Three segmentation models are trained separately on the CTB, MSR, and PKU datasets. The results from three models are directly combined as the MWS results[5].
3. **Joint**: We use the CTB, MSR, and PKU training sets to jointly train a segmentation model

[4] https://huggingface.co/bert-large-uncased

[5] Conflicting words are all included in the final results.

Dataset		Annotation	#Sents	#Words	OOV(%)
Train	CTB	SWS	16,091	437,991	-
	MSR	SWS	78,226	2,121,758	-
	PKU	SWS	46,815	1,097,839	-
	Pseudo	MWS	138,628	4,127,461	-
Dev	NEWS	MWS	1,000	31,477	4.69
Test	NEWS	MWS	2,000	63,108	4.96
	BAIKE	MWS	6,320	14,450	40.71

Table 3: Data statistics in our experiments. Pseudo refers to automatically generated pseudo data [6]. SWS and MWS stand for single-granularity labels and multi-granularity labels, respectively.

with a MTL method as describe in Section 4. Still, the results are directly used as the MWS results.

4. **Joint+Vote**: Similar to **Joint**, but the conflicts in the ouputs are resolved through a voting mechanism.
5. **Joint+CKY**: Similar to **Joint**, but we employ the CKY decoding to find the optimal conflict resolution.

6.2 Main Result

Table 4 compares various methods on the NEWS-test and BAIKE-test data.

Comparison with baselines. We first compare our method with single-task learning method (**Single**) and the tree parsing method (**Tree-based**) on NEWS-test and BAIKE-test datasets. We observe that **Single** achieves relatively high recall compared to other methods, but its precision is very low due to its disregard for connections among different heterogeneous SWS data. The **Tree-based** model achieves relatively high precision at the expense of a lower recall rate. In contrast, the proposed method (**Joint**) demonstrates significant enhancements on both the NEWS-test and BAIKE-test datasets, with F1 score improvements of 0.2 and 3.12 respectively compared to these two baseline methods. This underscores the suitability of our method for MWS tasks and its effectiveness in domain transfer.

Notably, **Tree-based** method only achieves performance similar to **Single** method when BERT is used, highlighting the drawbacks of utilizing pseudo MWS data as.

Impact of conflict resolution. We further investigate the impact the conflict resolution strategy.[7] Compared to **Joint**, which simply overlooks conflicts, both strategies show enhancements in performance. in particular, when utilizing BERT, our method demonstrates F1 score improvements of 0.63 and 1.68 on NEWS-test and BAIKE-test datasets, respectively, whereas **Joint+Vote** achieves F1 score improvements of 0.36 and 0.23. This highlights the advantageous nature of conflict resolution in the MWS task. Additionally, our proposed CKY decoding module shows more substantial improvements compared to the voting method, as the voting method struggles to resolve ties when options receive an equal number of votes. Finally, our method (**Joint+CKY**) outperforms the current SOTA model (**Tree-based**) by achieving improvements of 0.83 and 4.8 on the two test datasets.

Model	NEWS-test			BAIKE-test		
	P	R	F1	P	R	F1
BiLSTM						
Tree-based	**95.24**	90.59	92.86	**48.39**	43.30	40.59
Single	87.16	**93.95**	90.43	38.21	49.87	**46.94**
Joint	93.56	92.64	93.10	38.93	51.58	44.37
Joint+Vote	93.04	93.78	93.40	39.02	51.90	45.46
Joint+CKY	93.73	93.45	**93.59**	40.53	**52.93**	45.91
BERT						
Tree-based	94.69	92.05	93.36	56.17	63.68	59.93
Single	92.49	94.08	93.28	52.40	75.87	61.99
Joint	94.05	93.07	93.56	54.72	74.37	63.05
Joint+Vote	93.61	**94.25**	93.92	55.70	73.26	63.28
Joint+CKY	**95.26**	93.14	**94.19**	58.01	73.20	64.73
adding BAIKE	94.76	93.50	94.13	**60.74**	**74.60**	**66.96**

Table 4: The performance of different methods on in-domain NEWS-test and cross-domain BAIKE-test. adding BAIKE used 3 million BAIKE training data with marginal probabilities distributed between 0.4 and 1.0.

Utilization of naturally annotated data. We delve deeper into the efficacy of employing naturally annotated data from Baidu Baike for data augmentation. The results presented in Table 4 indicate that our data augmentation approach does not affect in-domain outcomes but leads to enhancements in the cross-domain BAIKE-test results, with improvements of 2.73 in precision, 1.4 in recall, and 2.23 in F1 score compared to **Joint+CKY**. In the subsequent section, we will conduct a more detailed analysis of the influence of BAIKE data on model performance.

[7] According to our statistical results, there are 1.7% conflicting words in NEWS-test.

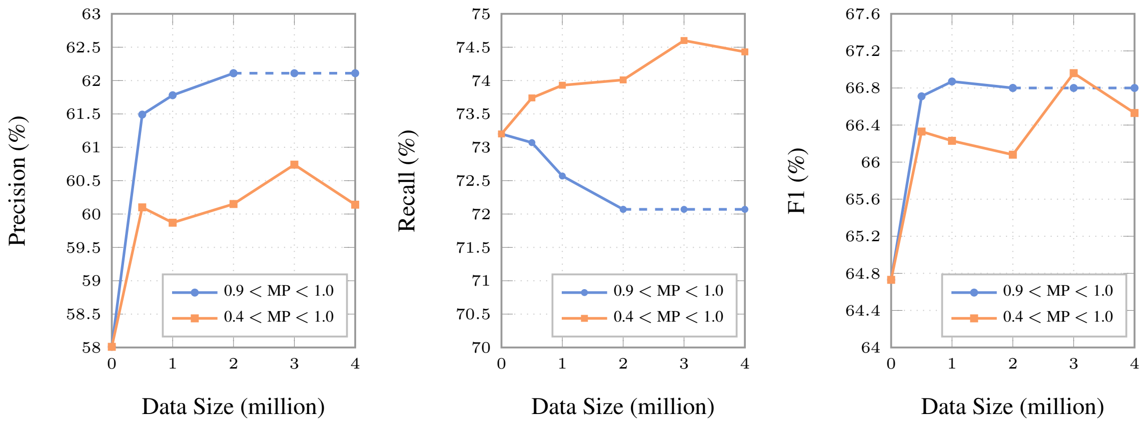

Figure 3: The performance on BAIKE-test when training examples of different scales and different probability intervals are added to model training, using precision, recall, and F1 score as reference metrics. The dashed line indicates the absence of actual values due to the limited number of data.

6.3 Analysis

Table 2 presents the statistics of the 5.2 million high-quality training examples selected from the 12 million BAIKE data [8]. We can identify two significant features: (1) Sentences with high probabilities have the highest quality and the fewest conflicts. (2) Segments with low probabilities exhibit a greater increase in the number of spans after being re-segmented.

We conducted extensive experiments by varying the data scale and source [9] to observe their impact on model performance.

Influence of the amount of data. Figure 3 indicates that as the scale of additional training data increases, the model's performance on the BAIKE-test improves. We can observed, as the data scale increases, both probability intervals contribute to improvements in the F1 score, with the maximum improvement being 2.23.

Performance of different marginal probability intervals. The results in Figure 3 indicate that sentences with high marginal probabilities notably enhance precision, while those with low marginal probabilities significantly improve recall. We observe that incorporating sentences with high marginal probabilities decreases recall. These sentences lack diversity in the overall data distribution, reinforcing the model's biases and reducing its robustness. Conversely, sentences with low marginal probabilities, as shown in Table 2, are more diverse and contain richer lexical information, thereby enhancing the model's generalization ability. However, from the distribution of F1 scores, both types of data contribute to the improvement in F1, indicating the effectiveness of our data augmentation method.

7 Conclusion

This work advances the state-of-the-art (SOTA) in MWS research through three key contributions. First, we apply span-based CWS methods to the MWS task, evaluating our model on both in-domain NEWS-test data and cross-domain BAIKE-test data. Second, we introduce the CKY decoding algorithm to resolve segmentation conflicts. Finally, we derive a substantial number of high-quality training examples from Baidu Baike texts by filtering based on marginal probabilities and employing a local loss function to enhance the model's performance on cross-domain test data. Extensive experiments demonstrate that: (1) integrating the span-based segmentation model with the CKY decoding algorithm for conflict resolution significantly enhances model performance; (2) filtering high-quality training examples based on marginal probabilities effectively facilitates domain transfer; and (3) our method improves the F1 score by 0.83 on the NEWS-test and by 7.03 on the BAIKE-test compared to the current SOTA MWS model.

[8] We collect 12 million sentences with natural annotation from Baidu Baike website: https://baike.baidu.com/

[9] Segment the sentence into multiple intervals based on marginal probabilities.

Limitations

While our approach demonstrates significant advancements over the current (SOTA) model across both in-domain and cross-domain test sets, particularly through effective data augmentation on the cross-domain BAIKE-test, there remains substantial room for further enhancement.

On the one hand, we identified incomplete labeling in the BAIKE-test dataset, where annotated segments lack a cohesive hierarchical structure in their labels. Given constraints on time and the extensive workload associated with re-annotation, we are presently unable to address these issues.

On the other hand, the availability of only one development set and two test sets for MWS limits our ability to comprehensively validate the superiority of our method across diverse domains. Lastly, computational resource constraints prevented us from utilizing larger-scale datasets for augmenting our data or exploring additional patterns effectively.

References

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial Multi-Criteria Learning for Chinese Word Segmentation. In *Proceedings of ACL*, pages 1193–1203, Vancouver, Canada.

Tzu Hsuan Chou, Chun-Yi Lin, and Hung-Yu Kao. 2023. Advancing multi-criteria Chinese word segmentation through criterion classification and denoising. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6460–6476, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Chenhui Dou, Chen Gong, Zhenghua Li, Zhefeng Wang, Baoxing Huai, and Min Zhang. 2024. Improving Chinese Named Entity Recognition with Multi-grained Words and Part-of-Speech Tags via Joint Modeling. In *Proceedings LREC-COLING*, pages 8732–8742.

Chen Gong, Zhenghua Li, Min Zhang, and Xinzhou Jiang. 2017. Multi-grained Chinese word segmentation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 692–703, Copenhagen, Denmark. Association for Computational Linguistics.

Chen Gong, Zhenghua Li, Bowei Zou, and Min Zhang. 2020. Multi-grained Chinese word segmentation with weakly labeled data. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2026–2036, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jingjing Gong, Xinchi Chen, Tao Gui, and Xipeng Qiu. 2019. Switch-LSTMs for Multi-Criteria Chinese Word Segmentation. In *Proceedings of AAAI*, pages 6457–6464.

Chang-Ning Huang, Yumei Li, and Xiaodan Zhu. 2006. Tokenization guidelines of chinese text (v5. 0, in chinese). *Microsoft Research Asia*.

Wenbin Jiang, Meng Sun, Yajuan Lü, Yating Yang, and Qun Liu. 2013. Discriminative Learning with Natural Annotations: Word Segmentation as a Case Study. In *Proceedings of ACL*, pages 761–769, Sofia, Bulgaria.

Yuanyuan Jin, Shiyu Tao, Qi Liu, and Xiaodong Liu. 2022. A BiLSTM-CRF Based Approach to Word Segmentation in Chinese. In *Proceedings of DASC*, pages 1–4, Falerna, Italy.

Tadao Kasami. 1965. An Efficient Recognition and Syntax-Analysis Algorithm for Context-Free Languages.

Hsiu-Wen Li, Ying-Jia Lin, Yi-Ting Li, Chun Lin, and Hung-Yu Kao. 2023. Improved Unsupervised Chinese Word Segmentation Using Pre-trained Knowledge and Pseudo-labeling Transfer. In *Proceedings of EMNLP*, pages 9109–9118.

Wei Li, Yuhan Song, Qi Su, and Yanqiu Shao. 2022. Unsupervised Chinese word segmentation with BERT oriented probing and transformation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3935–3940, Dublin, Ireland. Association for Computational Linguistics.

Yijia Liu, Wanxiang Che, Jiang Guo, Bing Qin, and Ting Liu. 2016. Exploring segment representations for neural segmentation models. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*.

Yijia Liu, Yue Zhang, Wanxiang Che, Ting Liu, and Fan Wu. 2014. Domain adaptation for CRF-based Chinese word segmentation using free annotations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 864–874, Doha, Qatar. Association for Computational Linguistics.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 562–568, Geneva, Switzerland. COLING.

Xipeng Qiu, Hengzhi Pei, Hang Yan, and Xuanjing Huang. 2020. A Concise Model for Multi-Criteria Chinese Word Segmentation with Transformer Encoder. In *Findings of EMNLP*, pages 2887–2897, Online.

Sriparna Saha and Asif Ekbal. 2013. Combining multiple classifiers using vote based classifier ensemble technique for named entity recognition. *DKE*, pages 15–39.

Xuewen Shi, Heyan Huang, Ping Jian, Yuhang Guo, Xiaochi Wei, and Yi-Kun Tang. 2019. Neural Chinese Word Segmentation as Sequence to Sequence Translation. *ArXiv preprint*, abs/1911.12982.

Youhyun Shin and Sang-goo Lee. 2020. Learning Context Using Segment-Level LSTM for Neural Sequence Labeling. *TASLP*, pages 105–115.

Charles Sutton, Andrew McCallum, and Khashayar Rohanimanesh. 2007. Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data.

Zhicheng Wang, Tianyu Shi, and Cong Liu. 2022. Joint Chinese Word Segmentation and Span-based Constituency Parsing. *ArXiv preprint*, abs/2211.01638.

Shiting Xu. 2024. BED: Chinese Word Segmentation Model Based on Boundary-Enhanced Decoder. In *Proceedings of CACML*, pages 263–270.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 8, Number 1, February 2003: Special Issue on Word Formation and Chinese Language Processing*, pages 29–48.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, pages 207–238.

Haiqin Yang. 2019. BERT Meets Chinese Word Segmentation. *ArXiv preprint*, abs/1909.09292.

Daniel H. Younger. 1967. Recognition and Parsing of Context-Free Languages in Time n^3. *Information and Control*, pages 189–208.

Shiwen Yu and Xuefeng Zhu. 1998. Dictionary of Modern Chinese Grammar Information.

Yu Zhang, Houquan Zhou, and Zhenghua Li. 2020. Fast and accurate neural CRF constituency parsing. *ArXiv preprint*, abs/2008.03736.

Hai Zhao, Deng Cai, Yang Xin, Yuzhu Wang, and Zhongye Jia. 2017. A Hybrid Model for Chinese Spelling Check. *TALLIP*, pages 21:1–21:22.

Lujun Zhao, Qi Zhang, Peng Wang, and Xiaoyu Liu. 2018a. Neural Networks Incorporating Unlabeled and Partially-labeled Data for Cross-domain Chinese Word Segmentation. In *Proceedings of IJCAI*, pages 4602–4608.

Yue Zhao, Hang Li, Shoulin Yin, and Yang Sun. 2018b. A New Chinese Word Segmentation Method Based on Maximum Matching. *JIHMSP*, pages 1528–1535.