Word Boundary Decision : An Efficient Approach for Low-Resource Word Segmentation

Yu Wang
The Hong Kong Polytechnic University
Hong Kong, SAR, China
janet-yu.wang@connect.polyu.hk

Chu-Ren Huang
The Hong Kong Polytechnic University
Hong Kong, SAR, China
churen.huang@polyu.edu.hk

Abstract

Due to the limitation of data, low-resource word segmentation poses significant challenges for pre-trained language models, which struggle to process new knowledge beyond their training data. Instead of focusing on data augmentation or transfer representations, this paper proposes an efficient approach called Word Boundary Decision (WBD), which redefines word segmentation learning goals as segmentation behaviors rather than segmented units from the training data. The paper presents experiments across diverse datasets, including social media, medical, patent, Cantonese, and ancient Chinese text. In small sample tests, WBD enables models to achieve the same performance with substantially less training data—for example, requiring only 3K words to match baseline F_1 scores at 20K words for ancient Chinese, representing around 6.67 times less data. Through transfer learning experiments, WBD also significantly enhances the cross-domain performance of pre-trained language models. For instance, WBD increases F_1 scores by 2.48% and R_{OOV} by 2.28% for BERT on average. This paper is an initial attempt to enable models to process new knowledge beyond their training data through task formulation[1].

1 Introduction

Due to the limitation of data, low-resource word segmentation poses significant challenges for pre-trained language models, which struggle to process new knowledge beyond their training data (Roberts et al., 2020; Yin et al., 2023; Hedderich et al., 2021a). To alleviate the issue, many methods have been proposed to improve pre-trained language models' performance in low-resource settings, such as data augmentation (Ding et al., 2020; Feng et al., 2021), distant and weak supervision (Hedderich et al., 2021b; Liang et al., 2020), cross-lingual projection (Cotterell and Duh, 2024; Liu et al., 2021), transfer learning (Alyafeai et al., 2020; Raffel et al., 2020), etc. These technologies aim to generate additional labeled data to extend the task-specific data or transfer learned representations from high-resource to low-resource domains to reduce the need for data. For example, Xing et al. (2018) propose an adaptive multi-task transfer learning approach to avoid the high annotation cost for collecting large scale word segmentation data for medical domain. In a similar vein, Ye et al. (2019) use a semi-supervised approach to improve word segmentation performance in novels, medicine, and patent cross-domain tasks. Additionally, Shen et al. (2022) use a data augmentation method to generate additional data for ancient Chinese word segmentation tasks. These research efforts achieve promising results by augmenting or leveraging limited available data.

However, the fundamental issue persists. When encountering word patterns that not shown in the training data, the performance drops significantly. Examples in low-resource languages include special expressions such as "戇居" (stupid) and "梗係" (surely) in Cantonese, as well as compound words that should be separated into multiple words in the training data but have been used as single words in specific text, such as "小九九" (literally: small nine nine; new meaning: trick) , highlighting the need for processing new knowledge, which pre-trained language models currently lack, leading to sub-optimal performance.

To address these limitations, taking Chinese word segmentation (CWS) in low resource as topic, this paper proposes an efficient word segmentation approach for pre-trained language models called word boundary decision (WBD). The core innovation of this method lies in:

Redefining the way pre-trained language models acquire "word segmentation" knowledge, transfer-

[1] Data: https://github.com/LANGUAGE-UNDERSTANDING/Word-Boundary-Decision-An-Efficient-Approach-for-Low-Resource-Word-Segmentation

ring the learning goal from learning instances to learning behaviors.

Departing from the conventional approaches of augmenting or leveraging data to combat low-resource challenges, this method tackles the problem from the task formulation level, enabling models to learn more knowledge by simpler design. Notably, the method can be combined with other methods for enhancing low-resource performance, such as transfer learning and data augmentation, offering a synergistic effect.

The main contributions are:

- We combined the formulation of Huang et al. (2007) with modern deep learning techniques and introduced an efficient approach, Word Boundary Decision (WBD) for low-resource scenarios, enabling models to achieve the same performance with substantially less training data – for example, requiring only 3K words to match baseline F_1 scores at 20K words for ancient Chinese, around 6.67 times less.
- Our WBD significantly improves transfer learning performance across various cross-domain sets, with F_1 scores increasing by 2.48%-10.46% and R_{oov} by 0.44%-5.26% for BERT and RoBERTa.
- To our knowledge, we are the first to test the robustness of models by checking the size of the required training dataset, which is an essential issue in low-resource areas.
- To our knowledge, we are the first to address the low-resource word segmentation issue from a task formulation perspective, redefining the training process to reduce the mimic phenomenon and enhance models' ability to process new knowledge beyond their training data.

2 Word Boundary Decision

2.1 Current Character-tagging Approach

In the era of pre-trained language models, the most dominant approach for word segmentation is the character-tagging approach. This approach treats word segmentation as a sequence labeling problem (Xue, 2003). For an input text sequence, the program annotates each character from left to right with corresponding labels, and then segments the text into separate words based on these labels. The most popular labeling tag set is T = B, M, E, S. This labeling is inspired by the classic BIO (Begin, Inside, Outside) scheme in the information extraction field, annotating characters as B (Begin, word beginning), M (Middle, word middle), E (End, word end), and S (Single, single-character word). After labeling, the program segments the text at the characters labeled as "E" (word end) or "S" (single-character word), thereby obtaining the corresponding word sequence. The goal of the character-tagging approach is to learn from the segmented units of the training data.

However, word segmentation aims to provide an appropriate separation between characters in a string without delimiters, for example, transforming "苹果和梨" (appleandpear) into "苹果/和/梨" (apple/and/pear) by providing a "/". It involves only one piece of information: whether to segment or not. On the other hand, the essence of character-tagging approach like {B, M, E, S} is to classify each character and determine its position within a word, and then convert this context-dependent information (word beginning, word middle, word end, etc.) into word boundary information (segment/not segment). This approach, which uses multi-class character classification information for single-class word delimiter recognition, introduces redundant information for the word segmentation task.

2.2 Word Boundary Decision Approach

Based on Huang et al. (2007), Li and Huang (2009), Huang and Xue (2012), this paper proposes a different perspective on word segmentation for pre-trained language models called word boundary decision (WBD). Instead of treating it as a character-tagging task, this approach views word segmentation as a word boundary decision process. The goal is to determine whether the boundary between characters is a word boundary. We formally represent a text segment as:

$$C_1, I_1, C_2, I_2, ..., C_i, I_i, ..., C_{n-1}, I_{n-1}, C_n$$

Where C_i represents a Chinese character, and I_i represents the boundary between characters C_i and C_{i+1} . In Chinese text, these character boundaries do not explicitly indicate whether they are word boundaries. We define that if a character boundary is a word boundary, it is denoted as $I_i = 1$, otherwise $I_i = 0$. The program segments the text based on the word boundary labels $I_i = 1$, completing the word segmentation task.

Comparing these two approaches, the character-tagging approach takes classifying characters as the target, designed to learn from the segmented units of the training data, using multi-class character classification information for single-class word delimiter recognition. This introduces redundant information, increasing the likelihood of repeating the same mistakes found in the training data and making it challenging to learn new knowledge, especially in low-resource scenarios.

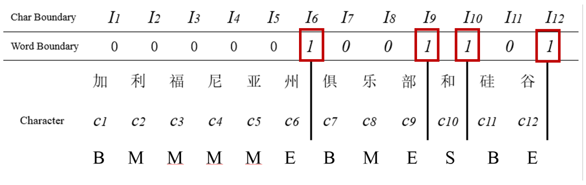

Char Boundary I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12
Word Boundary 0 0 0 0 0 1 0 0 1 1 0 1
加 利 福 尼 亚 州 俱 乐 部 和 硅 谷
Character c1 c2 c3 c4 c5 c6 c7 c8 c9 c10 c11 c12
B M M M M E B M E S B E

Figure 1: Examples of Word Boundary Decision (WBD) segmentation

In contrast, the WBD approach takes boundaries as the target, simplifying word segmentation into a binary decision for a single unit: whether a boundary is a word boundary or not. WBD learns from the segmentation behavior of the training data and does not involve the excessive information of segmented units. Hence, it is less likely to be misled by the training data and can better capture low-resource language-specific characteristics, exhibiting excellent robustness and generalization capabilities.

3 Experimental Setup

The experiment consists of two parts: small sample testing and transfer learning testing to evaluate performance of WBD in low-resource scenarios.

The experiments take PKU dataset from SIGHAN 2005 bakeoff (Emerson, 2005) as the training set and test the performance of WBD in pre-trained language models: BERT (Devlin et al., 2019) , and RoBERTa (Liu et al., 2019) on five open-source CWS datasets, ranging from different domains, time periods, and dialect variants, including social media text WEIBO (Qiu et al., 2016), medical text AMTTL (Xing et al., 2018), patent text PT (Ye et al., 2019), ancient Chinese EvaHan (Li et al., 2022), and Cantonese HKCC (Luke and Wong, 2015). Statistics of datasets are shown in Table 1.

3.1 Pre-processing

Pre-processing such as substituting digits, English letters, Chinese idioms, and long words with unique symbols are commonly employed to enhance the performance of CWS models (Huang et al., 2020a; Ke et al., 2021a). However, in our experiment, we refrain from using such techniques for fair comparison, focusing solely on the potential improvements offered by the WBD.

DATASET	PKU		WEIBO		AMTTL	
	Train	Test	Train	Test	Train	Test
WORD	1110K	104K	421K	44K	45K	13K
CHAR	1826K	173K	689K	73K	73K	21K
WORD TYPE	55K	13K	43K	11K	6K	3K
CHAR TYPE	5K	3K	4K	3K	2K	1K
WORD LENGTH	1.65	1.65	1.64	1.68	1.61	1.62

DATASET	PT		EvaHan		HKCC	
	Train	Test	Train	Test	Train	Test
WORD	481K	34K	166K	28K	83K	47K
CHAR	828K	56K	194K	33K	114K	65K
WORD TYPE	36K	4K	11K	3K	10K	4K
CHAR TYPE	3K	1K	3K	2K	2K	1K
WORD LENGTH	1.72	1.67	1.17	1.18	1.37	1.39

Table 1: Statistics of datasets

3.2 Evaluation

The number of labels used in WBD is different from character-tagging, so for evaluation we first align the labels before comparison. We uniformly convert the predicted results to {B, M, E, S}, and then perform the comparison[2]. For consistency, all segmentation results are automatically calculated with the script provided by previous research (Tian et al., 2020a, He et al., 2022a)[3].The metrics are F_1 scores and R_{OOV} (Recall of out-of-vocabulary).

$$P = \frac{Correct\ predicted\ words}{Total\ predicted\ words} \times 100\% \tag{1}$$

$$R = \frac{Correct\ predicted\ words}{Total\ actual\ words} \times 100\% \tag{2}$$

$$R_{OOV} = \frac{Correct\ predicted\ OOV\ words}{Total\ actual\ OOV words} \times 100\% \tag{3}$$

$$F_1 = \frac{2PR}{P+R} \tag{4}$$

3.3 Hyper-parameters

The experimental environment is Google Colab, with an NVIDIA® T4 GPU 16GB, and the deep learning framework is PyTorch. It took 40 hours

[2] The conversion method is as follows: first, we segment the predicted results based on the predicted labels to generate a text file with words separated by spaces. Then, we use the script to annotate the text with {B, M, E, S}, generating the {B, M, E, S} results.

[3] Examples: https://github.com/SVAIGBA/WMSeg/tree/master or https://github.com/Anzi20/WeiDC/blob/main/evaluate.py

on the GPU to conduct all experiments. It's worth noting that a single training process is not time-consuming, ranging from 2 minutes to 30 minutes, depending on the size of the training data. During training, the training set is divided into 80% for training and 20% for validation. The hyper-parameters settings used in this paper are shown in Table 2. For ease of comparison, the parameters remain unchanged across all experiments.

H-PARAM	VALUE
Max input sequence length	256
Learning Rate	2e-5
Batch size	32
Optimizer	Adam
Loss function	Cross-entropy loss function

Table 2: Hyper-parameters

4 Prior Experiment

4.1 Comparison with State-of-the-Art Models in High-Resource Settings

Prior to assessing the impact of WBD in low-resource scenarios, it is essential to evaluate its fundamental performance in high-resource settings. If the performance of WBD is only satisfactory in low-resource settings, the applicability of this approach would be constrained. Results in Table 3 shows that in golden SIGHAN 2005 datasets, without fine-tuning the parameters, our WBD's performance is close to the state-of-the-art record. For R_{OOV} metric on the AS and CITYU datasets, WBD even achieves new best performance, surpassing previous state-of-the-art methods.

Model	MSR		PKU		AS		CITYU	
	F_1	R_{OOV}	F_1	R_{OOV}	F_1	R_{OOV}	F_1	R_{OOV}
Chen et al. (2017)	96.04	71.6	94.32	72.67	94.75	75.37	95.55	81.4
Ma et al. (2018)	98.1	80.0	96.1	78.8	96.2	70.7	97.2	87.5
Gong et al. (2019)	97.78	64.2	96.15	69.88	95.22	77.33	96.22	73.58
Qiu et al. (2020)	98.05	78.92	96.41	78.91	96.44	76.39	96.91	86.91
Duan and Zhao (2020)	97.6	-	95.5	-	95.7	-	95.4	-
Huang et al. (2020b)	97.9	84.0	96.7	81.6	96.7	77.3	97.6	90.1
Tian et al. (2020b)	98.4	84.87	96.53	85.36	96.62	79.64	97.93	90.15
Ke et al. (2021b)	98.50	83.03	96.92	80.90	97.01	80.89	98.20	90.66
Nguyen et al. (2021)	98.31	85.32	96.56	85.83	96.62	79.36	97.74	87.45
He et al. (2022b)	98.28	86.39	96.59	87.21	96.76	80.23	97.79	87.58
Our WBD(BERT)	98.16	84.98	96.45	83.28	96.60	**85.84**	97.90	**92.15**

Table 3: Comparison of different models on CWS

4.2 Comparison with Large Language Models in Low Resource Settings

Prior to experiments, it is necessary to test the performance of Large Language Models (LLMs) such as GPT-4.0 on CWS in low resource. If LLMs' performance exceeded pre-trained language models such as BERT, then there would be no need to use pre-trained language models for CWS in low-resource scenarios, nor discuss the impact of WBD on pre-trained language models.

We extracted 50 sentences from the HKCC test set, which is a Cantonese dataset (a low-resource Chinese dialect), and then input them into each model with the prompt: "Please segment the following sentences with spaces between words." The test results are shown in Table 4[4].

Model	GPT 4.0	ChatGPT	Claude-3-Sonnet	Jieba	BERT_PKU_WBD	BERT_WBD
F_1	63.64%	63.64%	64.19%	78.45%	80.14%	**93.19%**
R_{oov}	60.15%	60.15%	62.72%	65.19%	72.24%	**89.20%**

Table 4: CWS performance of LLMs and BERT WBD

The results above clearly show that for low-resource languages such as Cantonese, LLMs perform poorly, failing to adapt and capture features of the language. Segmentation tools like Jieba also failed to meet expectations. However, with training data, pre-trained language models get more promising results. Even when trained on the PKU dataset, which consists of simplified news articles, the performance of BERT with WBD can achieve 80.14% in F_1 and 72.24% in R_{oov} on Cantonese, a significantly higher performance than LLMs. Upon deeper analysis, we found that LLMs can rarely recognize Cantonese words, and most Cantonese-specific words are uniformly divided into single-character words. The result will not change significantly with few-shot support.

In conclusion, to improve word segmentation in low-resource settings, further research and exploration of pre-trained model's word segmentation methods are necessary.

5 Experimental Results

5.1 Results of Small Sample Testing

To assess the performance of WBD in low-resource environments, we conducted small sample experiments. We adopted a word-based sampling method to unify the amount of information. From each dataset, we sampled 3K, 4K, 5K, 6K, 9K, and 20K words as the training sets, while the test set remained the corresponding complete test set.

The results in Table 5 demonstrate that our WBD significantly enhances the learning effectiveness in low-resource scenarios. For instance, in the case of Cantonese, WBD improved F_1 by 3.24% - 8.25%, with an average improvement of 4.79%, and R_{oov}

[4] The tests were conducted in Apirl, 2024.

		3K	4K	5K	6K	9K	20K
		F1 SCORE					
	WBD	22.39%	25.31%	30.85%	35.69%	41.82%	64.30%
	BASE	1.70%	5.40%	15.76%	18.59%	37.29%	63.73%
WEIBO	**Diff**	**20.70%**	**19.91%**	**15.09%**	**17.10%**	**4.53%**	**0.56%**
		OOV RATE					
	WBD	20.59%	22.48%	28.49%	34.84%	41.07%	64.17%
	BASE	1.04%	3.23%	14.32%	14.82%	37.55%	63.22%
	Diff	**19.56%**	**19.24%**	**14.17%**	**20.01%**	**3.52%**	**0.95%**

		3K	4K	5K	6K	9K	20K
		F1 SCORE					
	WBD	46.58%	48.23%	47.60%	49.07%	48.77%	49.98%
	BASE	41.97%	42.69%	43.19%	41.35%	46.80%	46.78%
Medical	**Diff**	**4.60%**	**5.54%**	**4.41%**	**7.72%**	**1.96%**	**3.19%**
		OOV RATE					
	WBD	34.96%	37.36%	35.38%	37.66%	36.80%	38.43%
	BASE	34.76%	34.34%	33.86%	33.36%	37.93%	37.75%
	Diff	**0.20%**	**3.01%**	**1.52%**	**4.30%**	**-1.13%**	**0.69%**

		3K	4K	5K	6K	9K	20K
		F1 SCORE					
	WBD	73.77%	74.03%	74.63%	78.03%	78.06%	78.24%
Ancient Chinese	**BASE**	73.65%	73.45%	72.02%	73.84%	73.02%	72.44%
	Diff	**0.13%**	**0.58%**	**2.62%**	**4.20%**	**5.04%**	**5.80%**
		OOV RATE					
	WBD	63.83%	61.873%	59.82%	65.40%	65.31%	66.36%
	BASE	60.60%	59.09%	56.61%	58.67%	56.96%	55.82%
	Diff	**3.23%**	**0.73%**	**8.79%**	**6.64%**	**9.41%**	**10.54%**

		3K	4K	5K	6K	9K	20K
		F1 SCORE					
	WBD	56.15%	59.66%	66.35%	66.91%	75.98%	/
	BASE	47.90%	55.56%	61.80%	63.67%	72.17%	/
Cantonese	**Diff**	**8.25%**	**4.10%**	**4.55%**	**3.24%**	**3.81%**	/
		OOV RATE					
	WBD	18.15%	16.37%	25.98%	28.82%	49.97%	/
	BASE	11.06%	11.44%	11.70%	21.12%	39.10%	/
	Diff	**7.09%**	**4.92%**	**14.28%**	**7.70%**	**10.87%**	/

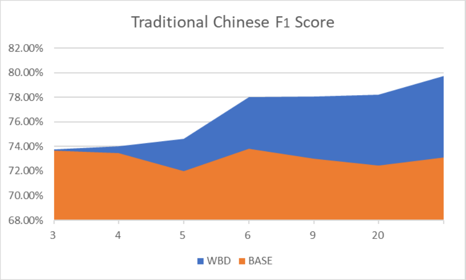
		3K	4K	5K	6K	9K	20K
		F1 SCORE					
	WBD	29.36%	31.40%	33.66%	35.02%	42.36%	56.44%
	BASE	15.45%	21.09%	28.67%	31.35%	41.98%	55.78%
Patent	**Diff**	**13.91%**	**10.31%**	**4.99%**	**3.67%**	**0.39%**	**0.66%**
		OOV RATE					
	WBD	24.89%	26.31%	27.22%	28.30%	34.84%	46.95%
	BASE	12.62%	18.81%	26.11%	26.50%	34.32%	47.12%
	Diff	**12.27%**	**7.50%**	**1.11%**	**1.80%**	**0.53%**	**-0.17%**

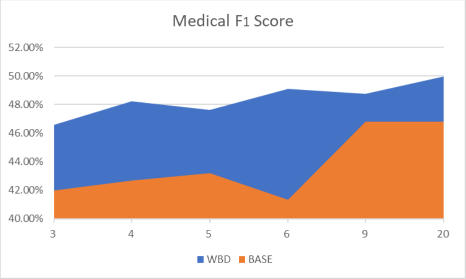
Table 5: Results of small sample testing

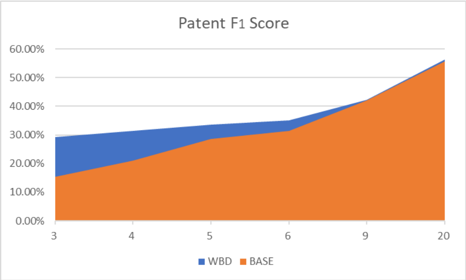
by 4.92%-10.87%, with an average improvement of 8.97%.

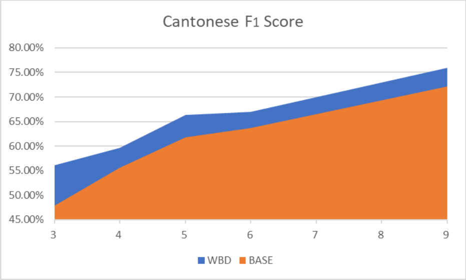
5.1.1 Required Data

Notably, WBD enabled the models to achieve the same CWS performance with significantly less required training data. As shown in Figure 2, for ancient Chinese, models with WBD (in blue) required only 3k training data to achieve the same F_1 as base models (in orange) with 20K training data, which is approximately 6.67 times less data. For medical text, models with WBD needed only 3k training data to achieve the same F_1 as base models with 9K training data, which is around 3 times less data.









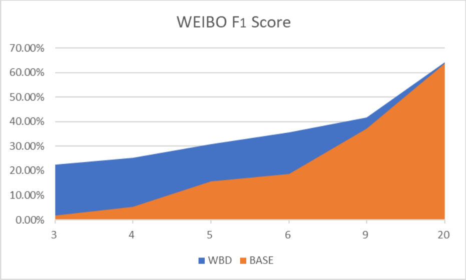


Figure 2: Area chart of F_1 score for small sample testing

These findings suggest that WBD substantially improves learning effectiveness, enabling models to capture new knowledge in low-resource, domain-specific datasets with much less required training data. This can greatly contribute to low-resource languages where training data is insufficient.

5.2 Results of Transfer Learning Testing

For transfer learning, we trained models on the PKU training set, which consists of simplified Chinese news from People's Daily, and evaluated their performance on five diverse cross-domain datasets: social media texts, medical texts, patent texts, ancient Chinese, and Cantonese. The results in Table 6 demonstrate that our WBD significantly enhances the transfer learning abilities of pre-trained language models. Specifically, WBD improved the average F_1 by 2.48% and R_{oov} by 2.28% for BERT, 2.30% and 2.85% respectively for RoBERTa.

Notably, WBD showed its most impressive improvements on the Cantonese dataset, with a remarkable 10.46% increase in F_1 and 5.26% in R_{oov} for RoBERTa. This could be due to the significant difference between Cantonese and the PKU dataset (simplified Chinese news from People's Daily). Cantonese is rich in traditional characters and single-character words (average word-length is 1.37), while most words in the PKU dataset are two or multi-character words (average word-length is 1.65). Conventional character-tagging approaches, which learn from segmented units in the training set, cannot capture the unique language characteristics, resulting in poor performance. However, WBD, which learns from boundary decision, a segmented behavior in the training set, demonstrates good adaptability to Cantonese, acquiring much more new language knowledge and showing remarkable improvement.

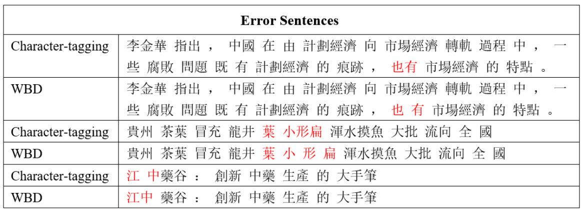
These findings clearly show that WBD is a powerful technique for boosting the cross-domain transfer capabilities of pre-trained language models, particularly in scenarios involving significant linguistic divergence from the training data.

6 Analysis and Discussion

We conducted an error analysis to explore why WBD enables pre-trained language models to achieve greater robustness and generalization capabilities, significantly improving performance in low-resource settings.

Comparing the segmented results by WBD and character-tagging, we found that there are mainly two types of errors that character-tagging models make but WBD models do not (examples are shown in Figure 3):

- Incorrectly combining frequently co-occurring individual words into one singer word; for example, mistakenly combine individual "也/ 有" into "也有".
- Ineffective recognition of less common collocations, such as mistakenly segmenting four words "葉/ 小/ 形/ 扁" as two words "葉/ 小形扁", single name entity "江中" as two words "江/ 中", and so on.



Error Sentences	
Character-tagging	李金華 指出 ， 中國 在 由 計劃經濟 向 市場經濟 轉軌 過程 中 ， 一些 腐敗 問題 既 有 計劃經濟 的 痕跡 ， 也有 市場經濟 的 特點 。
WBD	李金華 指出 ， 中國 在 由 計劃經濟 向 市場經濟 轉軌 過程 中 ， 一些 腐敗 問題 既 有 計劃經濟 的 痕跡 ， 也 有 市場經濟 的 特點 。
Character-tagging	貴州 茶葉 冒充 龍井 葉 小形扁 渾水摸魚 大批 流向 全 國
WBD	貴州 茶葉 冒充 龍井 葉 小 形 扁 渾水摸魚 大批 流向 全 國
Character-tagging	江 中藥谷 ： 創新 中藥 生產 的 大手筆
WBD	江中 藥谷 ： 創新 中藥 生產 的 大手筆

Figure 3: Examples of error sentences by Character-tagging

We conducted research on OOV (out-of-vocabulary) words in the output of transfer learning, where models trained on PKU were tested on various cross-domain datasets. The OOV words obtained by the WBD but not by the base models have very strong domain-specific features, such as Cantonese words like "戇居" (stupid) and "梗係" (surely), English expressions like "check" and "caibian3@peopledaily.com.cn", mixed-code words such as "b站" (short for "Bilibili", a website) and "A級" (A-level), as well as new meaning words like "老司机" (literally: old driver; new meaning: experienced person) and "二哈" (literally: two ha; new meaning: stupid Husky dog).

6.1 Explanation

To account for the phenomenon identified in error analysis, we need to first clearly define the difference between OOV and unknown words. OOV (out-of-vocabulary) words are defined according to an existing lexicon. Hence the term is more precise in describing a CWS that involves a word list. Unknown words are more broadly defined and could include OOV words. For clarity, we reserve this term to refer to words that are not recognized in the training data. In other words, they refer to words that should be recognized as segmentation units

MODELS	WEIBO		Medical		Patent		Ancient Chinese		Cantonese	
	F1	Roov	F1	Roov	F1	Roov	F1	Roov	F1	Roov
BERT_WBD	76.59%	52.67%	76.52%	44.30%	67.34%	45.87%	86.56%	76.69%	87.17%	71.12%
BERT_BASE	75.08%	49.24%	75.39%	42.25%	60.36%	42.97%	85.26%	74.42%	85.68%	70.40%
DIFFERENCE	**1.51%**	**3.43%**	**1.13%**	**2.06%**	**6.98%**	**2.90%**	**1.30%**	**2.27%**	**1.49%**	**0.72%**
RoBERTa_WBD	75.71%	52.16%	76.01%	44.98%	72.07%	53.53%	82.44%	72.69%	69.17%	64.22%
RoBERTa_BASE	75.22%	48.83%	75.20%	42.92%	71.62%	53.09%	81.31%	68.48%	58.72%	58.96%
DIFFERENCE	**0.49%**	**3.33%**	**0.81%**	**2.06%**	**0.46%**	**0.44%**	**1.13%**	**4.21%**	**10.46%**	**5.26%**

Table 6: Results of transfer learning testing

but are not segmented correctly in the training set, hence not attested and unknown.

Note that character-tagging models are trained based on the location and ordering of a character in a word (B, M, E, S). In other words, the accuracy of the information they provide depends on the training data's segmentation results. They are more likely to mimic the results of the training data. This is exactly what we see here. When training data incorrectly segments unknown words, a character-tagging model will most likely mirror that error.

WBD, on the other hand, classifies all between character blanks (potential word boundaries) and classifies them according to information obtained from various contexts defined by characters. That is, it is modeled in the context of characters, not words. The only segmentation-related information it uses from the training corpus is whether to segment or not in the context of that particular character string. It does not take into consideration the resulting words/segmentation units produced by the training data. Hence is it less likely to be misled by the training data's unknown words.

Based on the above, we can construct an explanation and argument why WBD will be likely to outperform a typical pre-trained based approach.

For segmentation tasks, it is reasonable to assume that typical pre-trained language models will be training based on the past results of segmentation units, although it may not be limited to the character location-in-a-word information as in the character-tagging model. It is expected to still have some over-fitting issues similar to other pre-trained language models based on previously segmented results.

WBD, on the other hand, only learns from the segmentation decision behavior on each boundary and does not learn from segmented units or involve the excessive information of these units. Therefore, it is not biased to make the same unknown word mistakes, making the model more robust and effective.

7 Conclusion

This paper proposes an efficient approach called Word Boundary Decision (WBD) for improving word segmentation performance of pre-trained language models, especially in low-resource scenarios. Unlike conventional character-tagging approaches that learn from the segmented units in the training data, WBD redefines word segmentation as a word boundary decision process, learning from the segmentation behaviors in the training data.

Through experiments on small sample testing and transfer learning across diverse datasets, the results demonstrate that WBD significantly enhances the learning effectiveness of pre-trained language models like BERT and RoBERTa. WBD achieves significant improvements in F_1 and R_{oov}, with the most remarkable gains observed for low-resource languages like Cantonese.

Notably, WBD enables the models to achieve the same performance with substantially less training data required compared to baselines (3K vs. 20K).

This method is an initial attempt to enable pre-trained language models to process new knowledge beyond their training data by task formulation.

8 Limitations

- **Lack of cross-lingual comparison.** Word segmentation tasks are not only applicable to Chinese, but also to other languages that lack explicit word delimiters, such as Japanese and Korean. There is a need to expand the scope of research to comprehensively compare and study the impact of WBD on word segmentation, leading to more robust conclusions.

- **Lack of exploration on synergistic effects.** WBD method can be combined with other methods such as transfer learning and data augmentation to form synergistic effect, which deserves further research.

- **Lack of more low-resourced cases.** For example , the minority language Yi, Vietnamese Chu Nom, and specific group scripts like Nüshu.

9 Ethics Statement

We affirm our commitment to contributing positively to society, prioritizing the avoidance of harm, and maintaining honesty and trustworthiness in our work. We do not anticipate any significant risks associated with our research. All experiments conducted in this study were based on publicly available datasets.

References

Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. 2020. A Survey on Transfer Learning in Natural Language Processing. *arXiv preprint*. ArXiv:2007.04239 [cs, stat].

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-criteria learning for chinese word segmentation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1193–1203. Association for Computational Linguistics.

Ryan Cotterell and Kevin Duh. 2024. Low-Resource Named Entity Recognition with Cross-Lingual, Character-Level Neural Conditional Random Fields. *arXiv preprint*. ArXiv:2404.09383 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint*. ArXiv:1810.04805 [cs].

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data Augmentation with a Generation Approach for Low-resource Tagging Tasks. *arXiv preprint*. ArXiv:2011.01549 [cs].

Sufeng Duan and Hai Zhao. 2020. Attention is all you need for chinese word segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 3862–3872, Online. Association for Computational Linguistics.

Thomas Emerson. 2005. The second international Chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A Survey of Data Augmentation Approaches for NLP. *arXiv preprint*. ArXiv:2105.03075 [cs].

Jingjing Gong, Xinchi Chen, Tao Gui, and Xipeng Qiu. 2019. Switch-lstms for multi-criteria chinese word segmentation. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6457–6464. AAAI Press.

Rian He, Shubin Cai, Zhong Ming, and Jialei Zhang. 2022a. Weighted self Distillation for Chinese word segmentation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1757–1770, Dublin, Ireland. Association for Computational Linguistics.

Rian He, Shubin Cai, Zhong Ming, and Jialei Zhang. 2022b. Weighted self distillation for Chinese word segmentation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1757–1770, Dublin, Ireland. Association for Computational Linguistics.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021a. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. *arXiv preprint*. ArXiv:2010.12309 [cs].

Michael A. Hedderich, Lukas Lange, and Dietrich Klakow. 2021b. ANEA: Distant Supervision for Low-Resource Named Entity Recognition. *arXiv preprint*. ArXiv:2102.13129 [cs].

Chu-Ren Huang and Nianwen Xue. 2012. Words without Boundaries: Computational Approaches to Chinese Word Segmentation. *Language and Linguistics Compass*, 6(8):494–505. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/lnc3.357.

Chu-Ren Huang, Petr Šimon, Shu-Kai Hsieh, and Laurent Prévot. 2007. Rethinking Chinese Word Segmentation: Tokenization, Character Classification, or Wordbreak Identification. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 69–72, Prague, Czech Republic. Association for Computational Linguistics.

Weipeng Huang, Xingyi Cheng, Kunlong Chen, Taifeng Wang, and Wei Chu. 2020a. Towards Fast and Accurate Neural Chinese Word Segmentation with Multi-Criteria Learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2062–2072, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Weipeng Huang, Xingyi Cheng, Kunlong Chen, Taifeng Wang, and Wei Chu. 2020b. Towards fast and accurate neural chinese word segmentation with multi-criteria learning. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 2062–2072, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Zhen Ke, Liang Shi, Songtao Sun, Erli Meng, Bin Wang, and Xipeng Qiu. 2021a. Pre-training with Meta Learning for Chinese Word Segmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5514–5523, Online. Association for Computational Linguistics.

Zhen Ke, Liang Shi, Songtao Sun, Erli Meng, Bin Wang, and Xipeng Qiu. 2021b. Pre-training with meta learning for chinese word segmentation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 5514–5523, Online. Association for Computational Linguistics.

Bin Li, Yiguo Yuan, Jingya Lu, Minxuan Feng, Chao Xu, Weiguang Qu, and Dongbo Wang. 2022. The First International Ancient Chinese Word Segmentation and POS Tagging Bakeoff: Overview of the EvaHan 2022 Evaluation Campaign. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 135–140, Marseille, France. European Language Resources Association.

Shoushan Li and Chu-Ren Huang. 2009. Word Boundary Decision with CRF for Chinese Word Segmentation. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation, Volume 2*, pages 726–732, Hong Kong. City University of Hong Kong.

Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1054–1064, Virtual Event CA USA. ACM.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint*. ArXiv:1907.11692 [cs].

K. K. Luke and M. L. Wong. 2015. The hong kong cantonese corpus: Design and uses. *Journal of Chinese Linguistics Monograph Series*, (25):312–333.

Ji Ma, Kuzman Ganchev, and David Weiss. 2018. State-of-the-art chinese word segmentation with bilstms. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4902–4908. Association for Computational Linguistics.

Duc-Vu Nguyen, Linh-Bao Vo, Dang Van Thin, and Ngan Luu-Thuy Nguyen. 2021. Span labeling approach for vietnamese and chinese word segmentation. In *PRICAI 2021: Trends in Artificial Intelligence - 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8-12, 2021, Proceedings, Part II*, volume 13032 of *Lecture Notes in Computer Science*, pages 244–258. Springer.

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 the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, pages 2887–2897, Online. Association for Computational Linguistics.

Xipeng Qiu, Peng Qian, and Zhan Shi. 2016. Overview of the NLPCC-ICCPOL 2016 Shared Task: Chinese Word Segmentation for Micro-Blog Texts. In Chin-Yew Lin, Nianwen Xue, Dongyan Zhao, Xuanjing Huang, and Yansong Feng, editors, *Natural Language Understanding and Intelligent Applications*, volume 10102, pages 901–906. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? *arXiv preprint*. ArXiv:2002.08910 [cs, stat].

Yutong Shen, Jiahuan Li, Shujian Huang, Yi Zhou, Xiaopeng Xie, and Qinxin Zhao. 2022. Data augmentation for low-resource word segmentation and pos tagging of ancient chinese texts. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 169–173.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020a. Improving Chinese Word Segmentation with Wordhood Memory Networks. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8274–8285, Online. Association for Computational Linguistics.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020b. Improving chinese word segmentation with wordhood memory networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8274–8285, Online. Association for Computational Linguistics.

Junjie Xing, Kenny Zhu, and Shaodian Zhang. 2018. Adaptive multi-task transfer learning for Chinese word segmentation in medical text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3619–3630.

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.

Yuxiao Ye, Yue Zhang, Weikang Li, Likun Qiu, and Jian Sun. 2019. Improving Cross-Domain Chinese Word Segmentation with Word Embeddings. In *Proceedings of the 2019 Conference of the North*, pages 2726–2735. ArXiv:1903.01698 [cs].

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do Large Language Models Know What They Don't Know? *arXiv preprint*. ArXiv:2305.18153 [cs].