Chinese Sequence Labeling with Semi-Supervised Boundary-Aware Language Model Pre-training

Longhui Zhang[1], **Dingkun Long, Meishan Zhang**[1]*
Yanzhao Zhang, Pengjun Xie, Min Zhang[1]
[1] Harbin Institute of Technology (Shenzhen), Shenzhen, China,
{longhuizhang97,longdingkun1993,zhangyanzhao00,xpjandy}@gmail.com
{zhangmeishan,zhangmin2021}@hit.edu.cn

Abstract

Chinese sequence labeling tasks are heavily reliant on accurate word boundary demarcation. Although current pre-trained language models (PLMs) have achieved substantial gains on these tasks, they rarely explicitly incorporate boundary information into the modeling process. An exception to this is BABERT (Jiang et al., 2022), which incorporates unsupervised statistical boundary information into Chinese BERT's pre-training objectives. Building upon this approach, we input supervised high-quality boundary information to enhance BABERT's learning, developing a semi-supervised boundary-aware PLM. To assess PLMs' ability to encode boundaries, we introduce a novel "Boundary Information Metric" that is both simple and effective. This metric allows comparison of different PLMs without task-specific fine-tuning. Experimental results on Chinese sequence labeling datasets demonstrate that the improved BABERT variant outperforms the vanilla version, not only on these tasks but also more broadly across a range of Chinese natural language understanding tasks. Additionally, our proposed metric offers a convenient and accurate means of evaluating PLMs' boundary awareness.

Keywords: Sequence Labeling, Boundary-Aware Language Model, Boundary Information Metric

1. Introduction

Sequence labeling is an important task in Chinese natural language processing (NLP), encompassing various tasks such as Chinese word segmentation (CWS), part-of-speech (POS) tagging, and named entity recognition (NER). These tasks inevitably rely on boundary identification among various grained semantic units, which are unavailable from the input Chinese sentences (Emerson, 2005a; Jin and Chen, 2008). There have been extensive studies that incorporate different types of boundary information into task-specific supervised machine learning models, i.e., subword-based models (Yang et al., 2019a; Li et al., 2021), lexicon-enhanced models (Zhang and Yang, 2018a; Diao et al., 2020a; Liu et al., 2021a). The results of these studies consistently demonstrate the high effectiveness of explicit boundary modeling in improving the performance of sequence labeling tasks in Chinese NLP.

Recently, Chinese PLMs, like BERT (Devlin et al., 2019), have shown significant success in various NLP tasks (Wei et al., 2020; Gao et al., 2021; Zhong and Chen, 2021), including sequence labeling (Yang et al., 2017; Jiang et al., 2021). BERT is efficient in capturing general semantic information, but it overlooks boundary information required for Chinese sequence labeling (Devlin et al., 2019; Diao et al., 2020b; Cui et al., 2021). Besides, substantial opportunities remain in research on the integration of boundary information into PLMs. By encoding boundary information into PLMs, we can potentially improve the performance of PLMs on various Chinese NLP tasks without task-specific optimizations, greatly benefiting the Chinese NLP community.

Chinese BABERT (Jiang et al., 2022) is one of the few exceptions that inject unsupervised statistical boundary information into vanilla BERT, resulting in considerable performance gains on Chinese sequence labeling tasks. Nevertheless, BABERT has a notable limitation: due to the long tail problem in calculating these unsupervised statistical signals, the statistical boundary information extracted from raw mining corpus could be unstable and low-quality. As such, there is an opportunity to further improve performance by exploring alternative sources of higher-quality boundary information more closely aligned with human intuitions.

Along the line of BABERT, in this work, we present Semi-BABERT, which supplements supervised boundary signals to BABERT. We construct a large-scale lexicon from open sources, which serves as a reliable resource for high-quality boundary information. To enhance BABERT pre-training, we design a span-based boundary recognition objective based on the boundary information extracted from the lexicon. Considering that boundary identification from a lexicon may be incomplete, we propose the utilization of Positive-Unlabeled learning (PU learning) (Li and Liu, 2005; Peng et al., 2019; Hu et al., 2021) to address this limitation and enable auto-complementation. Additionally, we introduce a practical metric to quantify the potential of Chinese PLMs in encoding boundaries without

* Corresponding author.

LREC-COLING 2024, pages 3179–3191

task-specific fine-tuning, which would be useful in swiftly assessing the adaptability of PLM to Chinese sequence labeling tasks.

We conduct comprehensive experiments to evaluate Semi-BABERT's capability in Chinese sequence labeling, covering various CWS, POS, and NER datasets. The results consistently demonstrate that Semi-BABERT improves performance across all benchmark datasets. In an effort to further substantiate the model's efficacy, we broaden the scope of our evaluation to other Chinese natural language processing tasks, such as text classification and machine reading comprehension. Similarly, we observe a marked increase in performance in these tasks, reaffirming the effectiveness of Semi-BABERT in a series of Chinese language understanding tasks. Finally, we present an in-depth analysis specifically focused on Chinese sequence labeling tasks. This analysis offers valuable insights into the impact of Semi-BABERT on different aspects of sequence labeling[1].

2. Method

In this section, we first briefly introduce BABERT, which serves as the foundation for our work. We then present our Semi-BABERT, which leverages a lexicon to incorporate large-scale, high-quality supervised boundary information into BABERT. Finally, a novel "Boundary Information Metric" is proposed, which can swiftly and efficiently quantify the model boundary awareness without task-specific fine-tuning.

2.1. BABERT

As depicted on the left side of Figure 1, BABERT (Jiang et al., 2022) incorporates unsupervised statistical boundary information into BERT's pre-training process, which consists of three steps.

1) The first step involves constructing a dictionary $\mathcal{N}$ to store the statistical boundary information. Each entry in the dictionary is a key-value pair, where the key represents an N-gram $g = \{c_1, \ldots, c_m\}$ derived from the corpus. The corresponding value comprises three statistical boundary measures: Pointwise Mutual Information (PMI), Left Entropy (LE), and Right Entropy (RE).

2) Utilizing the dictionary $\mathcal{N}$, a boundary-aware representation $\mathbf{E} = \{\mathbf{e_1}, ..., \mathbf{e_n}\}$ is constructed for each text $x = \{c_1, \ldots, c_n\}$ in the corpus. This representation incorporates the statistical boundary information.

3) The statistical boundary information is injected into the model using the unsupervised boundary-aware learning pre-training task (UBA task). Additionally, to capture general semantic information, BABERT incorporates BERT's Masked Language Modeling (MLM) task. Therefore, the overall loss function for BABERT is defined as follows:

$$\mathcal{L}_{\text{BABERT}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{UBA}} \quad (1)$$

The above description provides a high-level overview of BABERT to facilitate understanding of our proposed method. For a more comprehensive understanding, one can refer to the original paper (Jiang et al., 2022).

2.2. Semi-BABERT

To enhance the boundary encoding capability of PLM, we introduce Semi-BABERT, a novel approach that incorporates supervised lexicon boundary information into BABERT through a pre-training task called supervised boundary recognition (SBR task). As depicted on the right side of Figure 1, the training of Semi-BABERT consists of several modules: data source, data processing, and a new pre-training objective.

Data Source Our data is derived from two key sources: a knowledge graph and a crowded corpus for pre-training. The knowledge graph provides supervised lexical boundary information via rule-based filtering, whereas the crowded corpus supplies high-quality text data filtered by a large language model (LLM). Specifically, we employ the OwnThink Knowledge Graph[2], which encompasses both entity and regular words, providing rich boundary cues. For the crowded corpus, in line with BABERT (Jiang et al., 2022), we compile a mixed corpus from Chinese Wikipedia[3] and Baidu Baike[4] for our pre-training. This corpus consists of 3 billion tokens and 62 million sentences.

Data Preprocessing Data preprocessing plays a crucial role in ensuring the quality of training data. Here we describe the two main techniques we employ for data preprocessing: rule-based lexicon filtering and LLM-based corpus filtering.

• *Rule-based Lexicon Filtering*: To ensure lexical quality, we implement rule-based lexicon filtering, applying constraints to remove undesirable words. Specifically, we limit the length of words to 2-4 characters, remove shorter words that are nested or overlapping with other words, and eliminate words that contain punctuation marks or English characters. After applying these filtering rules, we obtain a lexicon of 30 million words.

[1]The source code and pre-trained Semi-BABERT will be publicly released at http://github.com/modelscope/adaseq/examples/semibabert.

[2]https://www.ownthink.com

[3]https://zh.wikipedia.org/wiki/

[4]https://baike.baidu.com/

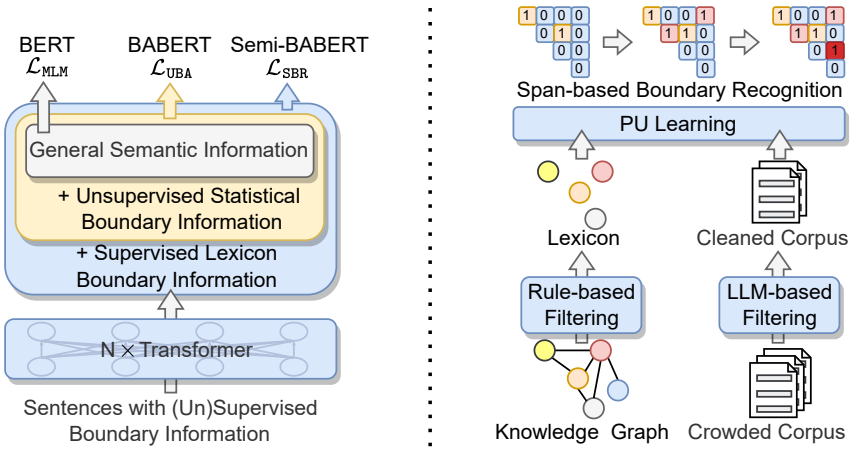


Figure 1: The relationship between BERT, BABERT and Semi-BABERT (left) and the supervised boundary recognition pre-training of Semi-BABERT (right).

• *LLM-based Corpus Filtering*: The quality of the pre-training corpus is crucial for the performance of PLMs (Liu et al., 2019). In our work, we leverage the power of LLMs for data cleaning to ensure the quality of the corpus used for pre-training. LLMs, thanks to their well pre-training and large-scale parameters, have demonstrated remarkable success in various natural language processing tasks (Liang et al., 2022) and are known to generate high-quality text (Chen et al., 2023). Therefore, we rely on the powerful generation capabilities of LLMs for data cleaning.

To evaluate the text quality in the crowded corpus, we introduce a task called "Text Quality Evaluation". The task prompt $\mathcal{P}$ is "请生成一个语法规范、表达准确、逻辑严谨的中文文本" (Please generate a grammatically correct, accurately expressed, and logically rigorous Chinese text). Given the task prompt $\mathcal{P}$ and a text $x = \{c_1, \ldots, c_n\}$, we calculate the text quality score $s(\mathcal{P}, x)$ using the following equation:

$$s(\mathcal{P}, x) = \frac{1}{n} \sum_i \log p\left(c_i \mid c_{<i}, \mathcal{P}\right) \quad (2)$$

$p\left(c_i \mid c_{<i}, \mathcal{P}\right)$ represents the generation probability of the character c_i conditioned on the prompt $\mathcal{P}$ and the previous text $c_{<i}$. We use the Qwen-7B (Bai et al., 2023) LLM for corpus filtering. To ensure efficiency, we first deduplicate the corpus and then evaluate the quality of all sentences in the corpus using Eq. 2. Finally, we remove the 10% of the corpus with the lowest score, thereby retaining 90%.

PU Learning PU learning algorithm (Li and Liu, 2005; Peng et al., 2019) trains a binary classifier f using only labeled positive examples $\mathcal{D}_p$ and a mixture of unlabeled positive and negative examples $\mathcal{D}_u$. This algorithm has shown success in distantly supervised NER (Peng et al., 2019). In PU learning, the loss function $\mathcal{L}(f)$ is defined based on $\mathcal{D}_p$ and $\mathcal{D}_u$ as follows:

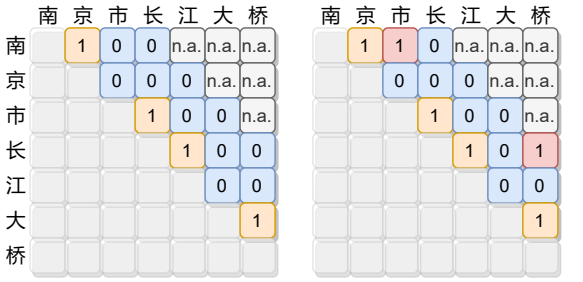
$$\mathcal{L}(f) = \gamma \pi_p \mathcal{L}_p^+(f) + \max\left\{0, \mathcal{L}_u^-(f) - \pi_p \mathcal{L}_p^-(f)\right\} \quad (3)$$

where $\mathcal{L}_{p/u}^{-/+}(f)$ represents the loss of an example $x \in \mathcal{D}_{p/u}$ conditioned on an assumed positive (denoted as "+") or negative (denoted as "-") label. The function $\mathcal{L}$ is a binary cross entropy loss function. π_p and γ are the hyperparameters of PU learning. π_p is the pre-estimated ratio of positive examples within $\mathcal{D}_u$. γ is the loss weight of $\mathcal{L}_p^+(f)$. The paper (Peng et al., 2019) proves a detailed proof of Eq. 3 in the PU learning algorithm.

Boundary Recognition Pre-training To address the instability and low quality of statistical boundary information in BABERT, we propose Semi-BABERT, which incorporates supervised lexicon boundary information. We introduce a supervised boundary recognition (SBR) task that aims to identify word boundaries in text based on the lexicon.

Traditionally, boundary recognition is regarded as a sequence labeling problem (Xu, 2003; Ding et al., 2020). However, this approach has two limitations. Firstly, it cannot handle nested boundaries, such as the distinction between "南京" (Nanjing) and "南京市" (Nanjing City). Secondly, supervision method heavily relies on lexical comprehensiveness, yet despite extensive size, the lexicon is unable to encompass all boundary information.

For the first drawback, we draw inspiration from research on structured information extraction (Yu et al., 2020; Ren et al., 2021) and adopt a span-based boundary recognition strategy instead of sequence labeling. This strategy involves training a binary classifier to determine whether N-grams in the text correspond to words. To address the second limitation, when dealing with an incomplete



(a) Initial boundary recognition (b) Final boundary recognition

Figure 2: Boundary identification results in the initial and final stages of the SBR task.

lexicon, we employ PU learning (Li and Liu, 2005; Peng et al., 2019; Hu et al., 2021) to estimate the model loss under the ideal complete lexicon, and gradually expand the lexicon to the ideal completeness through multiple iterations.

In our work, we focus on the span-based boundary recognition task (SBR). Given a lexicon and all N-grams present in the text, we consider the N-grams that exist in the lexicon as positive examples $\mathcal{D}_p$, while the remaining N-grams as unlabeled examples $\mathcal{D}_u$. To train a binary classifier f with PU learning, we formalize the PU learning algorithm using Eq. 3. During the training process, this algorithm helps us label the unlabeled N-grams in $\mathcal{D}_u$ as follows: If an unlabeled N-gram is predicted as a positive example by the classifier f consecutively for k times (k is set to 5), we add the N-gram to the lexicon in the next iteration. The overall loss $\mathcal{L}_{\text{SBR}}$ of SBR task and the classifier f are defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{SBR}} &= \mathcal{L}(f) \\ f((\mathbf{h}_i, \mathbf{h}_j)) &= \texttt{sigmoid}(W\,[\mathbf{h}_i; \mathbf{h}_j] + b) \end{aligned} \tag{4}$$

The span-based boundary recognition binary classifier f is trained to determine whether the N-gram is a correct boundary. It takes the PLM representations $\mathbf{h}_i$ and $\mathbf{h}_j$ corresponding to the left and right boundary characters of the N-gram $\{c_i, \ldots, c_j\}$ as input. The classifier predicts the probability that the N-gram is a word, using a sigmoid activation function applied to the linear transformation of the concatenation of $\mathbf{h}_i$ and $\mathbf{h}_j$ with weight matrix W and bias b.

In Figure 2, we demonstrate an example of the SBR task with PU learning. We consider the text "南京市长江大桥" (Nanjing Yangtze River Bridge). In the initial stage, only four words "南京" (Nanjing), "市长" (mayor), "长江" (Yangtze River), and "大桥" (Bridge) are included in the lexicon, as shown in Figure 2(a). However, with the help of PU learning and multiple iterations of the model, the boundary information of all N-grams such as "南京市" (Nanjing City) and "长江大桥" (Yangtze River Bridge) can be identified and added to the lexicon, as illustrated in Figure 2(b). It is important to note that we only consider N-grams $\{c_i, \ldots, c_j\}$ where $1 < j - i < 4$. Hence, the lower triangle area is ignored as $j - i \leq 1$, and certain parts of the upper triangle area do not require prediction as $j - i \geq 4$.

Pre-training Objective The architecture of Semi-BABERT, depicted on the left side of Figure 1, builds upon the foundations of BERT and BABERT. Therefore, apart from the span-based boundary recognition task, Semi-BABERT aligns with the training objectives of these base models. Consequently, the total pre-training loss for Semi-BABERT can be formulated as follows:

$$\mathcal{L}_{\text{Semi-BABERT}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{UBA}} + \mathcal{L}_{\text{SBR}} \tag{5}$$

2.3. Boundary Information Metric

Given the close relationship between the boundary awareness capability of PLMs and their performance on downstream Chinese sequence labeling tasks during fine-tuning, it is crucial to evaluate the boundary awareness capability of these PLMs. Previous research relied on downstream task performance as an evaluation metric (Liu et al., 2021a; Jiang et al., 2022). Although this evaluation metric is valid, it is unintuitive and resource-consuming. To address this issue and provide a more reasonable assessment for Chinese PLMs' boundary awareness, we introduce a novel evaluation metric called the "Boundary Information Metric" (BIM). This metric can evaluate PLMs' boundary recognition ability without task-specific fine-tuning.

We begin by assuming that: an ideal boundary-aware sentence representation should exhibit higher similarity between characters within words [5]. To quantify the boundary information, we consider a character c and define any other character within the same word as a positive sample c^+, while the character in a different word is considered to be a negative sample c^-. We calculate the similarity between c and c^+, denoted as $\texttt{SIM}_{pos}$. Similarly, the similarity between negative samples c and c^- is denoted as $\texttt{SIM}_{neg}$. The BIM is then calculated based on the difference between $\texttt{SIM}_{pos}$ and $\texttt{SIM}_{neg}$. We employ the cosine function to calculate the similarity between the vector representations $\mathbf{h}$, $\mathbf{h}^+$, and $\mathbf{h}^-$ corresponding to c, c^+, and c^-, respectively. A higher BIM indicates a stronger boundary awareness in the model. The final computation process of BIM can be formulated as:

$$\begin{aligned} \texttt{SIM}_{pos} &= \mathop{\mathbb{E}}_{(\mathbf{h},\mathbf{h}^+)\sim p_{\text{pos}}} \texttt{sim}\left(\mathbf{h}, \mathbf{h}^+\right) \\ \texttt{SIM}_{neg} &= \mathop{\mathbb{E}}_{(\mathbf{h},\mathbf{h}^-)\sim p_{\text{neg}}} \texttt{sim}\left(\mathbf{h}, \mathbf{h}^-\right) \\ \texttt{BIM} &= \texttt{SIM}_{pos} - \texttt{SIM}_{neg} \end{aligned} \tag{6}$$

[5] This assumption is empirically verified in Section 3.4

Intuitively, the similarity between character pairs not only depends on whether they are within the same word but also on their distance from the whole text. Characters farther apart tend to have lower similarity. To mitigate the influence of character distance on the BIM, we constrain the distance between the negative sample pair (c, c^-). Specifically, we make DIS(c, c^-) approximately equal to DIS(c, c^+), where DIS notes the distance between characters. Since c and c^+ are within the same word and are typically close to each other, we restrict the distance between c and c^-, i.e., the DIS(c, c^-) to a value less than a threshold L. In our works, we set L to 2. This constraint facilitates accurately quantify the boundary awareness of PLMs regardless of the character distance in the text.

3. Experiments

3.1. Datasets

In order to assess the effectiveness of Semi-BABERT for Chinese sequence labeling tasks, we conduct an extensive evaluation on a total of 13 diverse datasets. These datasets encompass a range of Chinese NLP tasks, including CWS, POS, and NER. These tasks are particularly relevant for evaluating the boundary awareness of PLMs, as they heavily rely on accurate boundary information.

For the CWS task, we utilize three datasets: CTB6 (Xue et al., 2005), MSRA, and PKU (Emerson, 2005b). For the POS task, we deploy three different datasets: CTB6 (Xue et al., 2005), UD1, and UD2 (Nivre et al., 2016; Shao et al., 2017). Lastly, for NER, we employ a total of seven datasets: Onto4 (Weischedel et al., 2011), Book (Jia et al., 2020), News (Jia et al., 2020), Finance (Jia et al., 2020), MSRA (Levow, 2006a), Resume (Yang et al., 2019b), and Weibo (Peng and Dredze, 2015, 2016).

3.2. Experimental Settings

We conduct pre-training of Semi-BABERT on a distributed setting using eight NVIDIA Tesla V100 GPUs, each equipped with 32GB memory. The hyperparameters and configurations of the baseline PLMs and Semi-BABERT are as follows:

Hyperparameters Following BABERT (Jiang et al., 2022), during pre-training, we use vanilla BERT to initialize the weights of Semi-BABERT. To accommodate different contexts, we pre-train two variations of Semi-BABERT, namely Semi-BABERT-base and Semi-BABERT-lite. Semi-BABERT-lite consists of 6 transformer layers, 8 self-attention heads, a hidden dimension of 512, and a total of 30 million parameters. In contrast, Semi-BABERT-base comprises 12 transformer layers, 12 self-attention heads, a hidden dimension of 768, and a total of 110 million parameters. For both variations, we adopt the same training hyperparameters: a batch size of 4096, a learning rate of 1e-4, a warmup ratio of 0.1, a maximum sentence length of 512, and a maximum N-gram length of 4. In the PU learning equation (Eq. 3), we set π_p to 0.2 and γ to 0.5.

Baselines To evaluate the effectiveness of Semi-BABERT, we compare its performance against several baseline models, including the following state-of-the-art approaches: BERT-lite (Devlin et al., 2019), BERT (Devlin et al., 2019), BERT-wwm (Cui et al., 2021), ERNIE-Baidu (Sun et al., 2019), ERNIE-Gram (Xiao et al., 2021), ZEN (Diao et al., 2020a), NEZHA (Wei et al., 2019), and BABERT (Jiang et al., 2022).

To ensure a fair and consistent comparison, we adopt the fine-tuning approach proposed by BABERT (Jiang et al., 2022), which incorporates PLMs with a conditional random field (CRF) layer for sequence labeling. During the inference stage, we utilize the Viterbi algorithm to generate the optimal label sequence. To mitigate the effects of randomness, we perform fine-tuning using five different random seeds and subsequently average the results. This process allows us to obtain robust and reliable performance estimates for Semi-BABERT and the baselines, enabling a comprehensive and unbiased comparison.

3.3. Main Results

In this section, we present the fine-tuning results of Semi-BABERT on 13 sequence labeling tasks, comparing it to various baselines. Table 1 summarizes the results, and we draw the following observations:

(1) Effectiveness of Semi-BABERT: Among models of the same scale, Semi-BABERT consistently achieves the highest average performance across all datasets. Specifically, Semi-BABERT-lite outperforms BERT-lite. When compared to the state-of-the-art BABERT, Semi-BABERT-base exhibits an average score increase of 0.8 (90.4-89.6) across all datasets. These results clearly demonstrate that Semi-BABERT surpasses PLMs of the same size, emphasizing the significance of supervised boundary information in achieving superior performance.

(2) Importance of boundary information: Apart from the scale of the training data, boundary information plays a crucial role in sequence labeling tasks. NEZHA, for instance, was pre-trained on three datasets (Chinese Wikipedia, Baidu Baike, and Chinese News) with a collective token count of 11 billion, four times larger than our dataset (Wei et al., 2019). Compared to NEZHA, the average score of Semi-BABERT with enhanced boundary information increases by 0.7 (90.3-89.6). Further-

Model	CWS			POS			NER							AVG
	CTB6	MSRA	PKU	CTB6	UD1	UD2	Onto4	Book	News	Finance	MSRA	Resume	Weibo	Score
BERT-wwm	97.4	98.3	96.5	94.8	95.5	95.4	80.9	76.2	79.3	85.0	95.7	95.8	68.6	89.2
ERNIE-Baidu	97.4	98.2	96.3	94.9	95.3	95.1	80.4	76.6	80.4	86.0	95.1	95.6	70.0	89.3
ERNIE-Gram	97.3	98.3	96.4	94.9	95.3	95.2	81.0	77.2	80.0	85.3	95.8	95.6	68.4	89.3
ZEN	97.3	98.3	96.5	94.8	95.5	95.5	80.1	75.7	80.2	85.0	95.2	95.4	66.7	88.9
NEZHA	**97.5**	**98.6**	96.7	95.0	95.6	95.5	81.7	77.0	79.8	85.2	**96.6**	95.7	70.3	89.6
BERT	97.3	98.2	96.3	94.7	95.0	94.9	81.0	76.1	79.2	85.3	95.8	95.6	69.6	89.2
BABERT	97.5	98.4	96.7	95.0	95.7	95.5	81.9	76.8	80.3	86.9	96.3	95.8	68.3	89.6
Ours	97.4	**98.6**	**96.8**	**95.2**	**95.7**	**95.6**	**82.2**	**80.7**	**81.9**	**87.1**	96.3	**96.0**	**71.0**	**90.3**
BERT-lite	97.0	98.1	96.1	94.4	93.7	93.3	77.2	76.0	79.1	83.9	93.9	95.4	64.4	87.9
Ours-lite	97.0	98.2	96.4	94.5	94.6	94.4	78.8	80.1	80.2	84.3	94.4	95.5	65.9	88.8
ChatGPT	94.3	92.9	93.1	88.6	92.0	92.1	69.4	70.9	80.4	79.3	90.1	95.7	70.1	85.3

Table 1: Fine-tuning results on Chinese sequence labeling tasks. We report the F1-score on the test set.

CWS	请对文本进行分词。输出形式为“单词1/单词2”。文本: [TEXT]。输出: Please segment the text into words. Output format is “word1/word2”. Text: [TEXT]. Output:
POS	请对下面的文本进行词性标注，其中词性列表为[Category List]。输出形式为“type1: word1; type2: word”。文本: [TEXT]。输出: Please provide part-of-speech tagging for the following text, where the tag list is [Category List]. Output format is “type1: word1; type2: word”. Text: [TEXT]. Output:
NER	请列出文本中所有符合下列类别的实体。输出形式: “类别1: 实体1; 类别2: 实体2;”。类别: [Category List] 文本: [TEXT]。输出: Please list all entities in the text that fit the following category. Output format is “type1: entity1; type2: entity2;”. Category: [Category List]. Text: [TEXT]. Output:

Table 2: ChatGPT prompts for three tasks.

more, despite the limited training data from Resume and Weibo (Liu et al., 2021a), Semi-BABERT still outperforms other models. These findings indicate that the inclusion of boundary information can compensate for the lack of training data.

(3) Importance of model size: The incorporation of high-quality boundary information enables the 6-layer Semi-BABERT-lite to surpass the 12-layer BERT on the Book and News datasets of NER task. However, BERT still maintains an average score 0.4 points higher than Semi-BABERT-lite (89.2-88.8), underscoring the continued significance of the size of PLMs, despite the gains made by Semi-BABERT-lite from incorporating boundary information.

In addition to the Chinese PLMs, LLMs have also achieved impressive performance in various tasks (Liang et al., 2022). ChatGPT [6] is the most representative among them. To evaluate its performance on Chinese sequence labeling tasks, we conduct further tests. Due to ChatGPT's closed-source setup and large-scale parameters, fine-tuning it becomes challenging. As a result, we only test ChatGPT in unsupervised scenarios. The prompts used for evaluation are provided in Table 2. The averages scores in Table 1 demonstrate that unsupervised ChatGPT remains less effective than supervised PLMs. Nevertheless, unsupervised ChatGPT exhibited noteworthy progress on certain datasets. Specifically, unsupervised ChatGPT outperforms supervised BERT on the News and Weibo datasets of the NER task. This outcome may be attributable to the limited scale of these datasets (Liu et al., 2021a), which likely provided inadequate training for the supervised PLMs.

[6] https://openai.com/blog/chatgpt

	PKU			Onto4		
	10	50	100	10	50	100
BERT-wwm	84.7	88.0	88.8	12.8	43.1	59.4
ERNIE	84.3	87.0	88.2	19.9	43.0	50.8
ERNIE-Gram	84.0	86.6	88.0	28.4	45.9	60.0
NEZHA	84.4	88.7	89.7	14.5	44.1	59.2
BERT	84.0	87.9	88.2	14.9	42.4	58.0
BABERT	84.7	89.5	90.0	32.1	46.6	60.6
Ours	**85.2**	**90.5**	**92.0**	**50.8**	**59.2**	**64.5**
BERT-lite	80.3	84.2	85.6	12.8	40.1	57.0
Ours-lite	84.8	89.6	91.8	47.6	57.1	59.7

Table 3: Few-shot results on PKU (CWS task) and Onto4 (NER task), using 10, 50, and 100 examples of the training data.

3.4. Analysis

In this section, we provide a comprehensive analysis of Semi-BABERT from four different perspectives: few-shot setting, probing, BIM evaluation, and case study.

Few-Shot PLMs have shown great performance in low-resource scenarios due to their extensive pre-training. To further investigate the capabilities of Semi-BABERT, we conduct fine-tuning experiments on various PLMs using 10, 50, and 100 randomly selected examples from the original training data of PKU (CWS) and Onto4 (NER) datasets.

	PKU	UD2	Onto4
BERT-wwm	88.8	55.3	32.5
ERNIE-Baidu	88.4	55.7	42.2
ERNIE-Gram	87.9	54.9	32.9
ZEN	88.7	55.6	31.8
NEZHA	88.7	55.9	38.3
BERT	87.4	54.7	31.2
BABERT	88.9	56.1	44.2
Ours	**89.2**	**56.5**	**45.7**

Table 4: The results of fine-tuning with frozen PLM on PKU (CWS), UD2 (POS) and Onto4 (NER).

Table 3 presents the results of these few-shot experiments, and it is evident that Semi-BABERT-base consistently outperforms various baselines across all few-shot settings. In the 10-shot setting of Onto4, Semi-BABERT-base achieves a remarkable score increase of 18.7 (50.8-32.1) points compared to BABERT, showcasing its effectiveness in low-resource scenarios. Additionally, the 6-layer Semi-BABERT-lite performs better than any other 12-layer PLMs, except in the 100-shot scenario of Onto4. These findings highlight the considerable performance gains attainable by effectively encoding boundary information, particularly in few-shot situations.

Probing The evaluation method based on full-parameter fine-tuning primarily assesses the performance of the fine-tuned models rather than the pre-trained ones. To gain further insights, we adopt a straightforward approach where the PLMs are frozen during fine-tuning, and only additional parameters, such as the CRF layer, are trained. Table 4 presents the results obtained using this fine-tuning method for various PLMs.

Interestingly, under the PLM frozen setting, Semi-BABERT demonstrates clear advantages over the full-parameter fine-tuning approach. Specifically, when comparing to BABERT in the full-parameter fine-tuning scenario (Table 1), Semi-BABERT achieves improvements of 0.1, 0.1, and 0.3 on the PKU, UD2, and Onto4 datasets, respectively. However, in the PLM frozen scenario (Table 4), these improvements are enhanced to 0.3, 0.4, and 1.5, respectively. This significant difference in performance highlights the effectiveness of Semi-BABERT's ability to enhance boundary awareness during pre-training.

BIM Evaluation In this subsection, we evaluate the effectiveness of the BIM, a boundary-aware quantification method that does not require task-specific fine-tuning. To assess the performance of BIM, we apply it to various PLMs. As BIM requires sentence segmentation, we conduct experiments on the CTB6 test set (Xue et al., 2005) of the Chinese Word Segmentation task, which provides high-quality word segmentation annotations.

	$\texttt{SIM}_{pos}$	$\texttt{SIM}_{neg}$	BIM
BERT-wwm	72.0	60.7	11.2
ERNIE-Baidu	78.1	64.5	13.6
ERNIE-Gram	86.2	72.6	13.6
ZEN	81.7	70.6	11.1
NEZHA	48.9	35.1	13.8
BERT	67.6	57.1	10.5
BABERT	65.2	51.2	14.0
Ours	62.5	47.3	**15.2**
BERT-lite	59.0	48.6	10.5
Ours-lite	57.8	43.6	14.1

Table 5: $\texttt{SIM}_{pos}$, $\texttt{SIM}_{neg}$ and BIM of various PLMs.

Text	黑手党2$_{\texttt{game}}$类似于游戏《侠盗飞车》$_{\texttt{game}}$，气氛有之前上映的黑帮传奇$_{\texttt{movie}}$的味道。 Mafia 2$_{\texttt{game}}$ is akin to Grand Theft Auto$_{\texttt{game}}$ with an ambiance reminiscent of the prior movie Gangster Legend$_{\texttt{movie}}$.
BERT	黑手党$_{\texttt{game}}$2类似于游戏《侠盗飞车》$_{\texttt{game}}$，气氛有之前上映的黑帮$_{\texttt{org}}$传奇的味道。
BABERT	黑手党2$_{\texttt{game}}$类似于游戏《侠盗飞车》$_{\texttt{game}}$，气氛有之前上映的黑帮$_{\texttt{org}}$传奇的味道。
Ours	黑手党2$_{\texttt{game}}$类似于游戏《侠盗飞车》$_{\texttt{game}}$，气氛有之前上映的黑帮传奇$_{\texttt{movie}}$的味道。

Table 6: Case study on NER task. Red (Blue) represents correct (incorrect) entities.

The results, presented in Table 5, demonstrate that for all models, $\texttt{SIM}_{pos}$ (the similarity between characters within words) is consistently higher than $\texttt{SIM}_{neg}$ (the similarity between characters across word boundaries). This observation validates BIM's underlying assumption that character representations within words tend to be more similar. Additionally, Semi-BABERT-base consistently achieves a higher BIM score compared to other PLMs, reaffirming its superior boundary awareness.

Two phenomena are surprising: (1) BERT-lite and BERT yield the same BIM score. (2) Semi-BABERT-lite achieves a higher BIM score than the larger-scale BABERT. The results suggest that the BIM is independent of the model scale, as it primarily measures boundary awareness of the model. Consequently, BIM serves as a more accurate quantification method for boundary information than task-specific fine-tuning.

Case Study To illustrate the clear advantages of Semi-BABERT over BERT and BABERT, we present a case study focusing on the NER task, as shown in Table 6. In this case, BERT, lacking explicit boundary information, fails to correctly identify the entities "黑手党2" (Mafia 2) and "黑帮传奇"

Model	TC						MRC			AVG
	AFQMC	TNEWS	IFLYTEK	OCNLI	WSC	CSL	CMRC	ChID	C3	Score
MacBERT	69.9	57.9	60.4	67.4	74.7	82.1	73.5	79.5	58.9	69.4
PERT	73.6	54.5	57.4	66.7	76.1	**82.8**	73.8	80.2	58.0	69.2
NEZHA	**73.5**	58.5	55.7	**69.0**	76.7	82.6	71.9	87.1	75.2	72.2
ERNIE-THU	72.9	56.6	59.3	68.0	75.8	82.4	73.0	80.2	56.3	69.4
ERNIE-Baidu	73.1	56.2	60.1	67.5	75.8	82.1	72.9	80.0	57.6	69.5
K-BERT	73.2	55.9	60.2	67.8	76.2	82.2	72.7	80.3	57.5	69.6
CKBERT	73.2	56.4	60.7	68.5	76.4	82.6	73.6	81.7	57.9	70.1
BERT	72.7	55.2	59.5	66.5	72.5	81.8	73.4	79.2	57.9	68.8
BABERT	71.1	57.1	60.0	65.6	73.1	80.2	71.3	83.1	68.1	70.0
Ours	73.1	**59.4**	**61.9**	68.5	**80.3**	81.1	**73.9**	**87.4**	**77.0**	**73.6**

Table 7: Performance of different PLMs on CLUE benchmarks.

(Gangster Legend). Although BABERT improves upon BERT by successfully identifying "黑手党2" (Mafia 2), it still falls short with an incomplete entity. Only Semi-BABERT accurately identifies all entities and their respective categories. This case study serves as a compelling demonstration of how high-quality boundary information can significantly enhance the performance of models on sequence labeling tasks.

3.5. Additional Results on CLUE

To assess the broader effectiveness of Semi-BABERT across various tasks, we conduct experiments on the widely-used CLUE benchmark (Xu et al., 2020a), specifically focusing on Chinese text classification (TC) and machine reading comprehension (MRC) tasks. The TC task consists of 6 datasets: AFQMC, TNEWS, IFLYTEK, OCNLI, WSC, and CSL. The MRC task comprises 3 datasets: CMRC, ChID, and C3. Following (Zhang et al., 2022), we use the fine-tuning code provided by CLUE benchmarks[7].

In addition to boundary information, the CLUE benchmarks also require models to possess significant knowledge, including structured relational information in a knowledge graph (Zhang et al., 2022). Therefore, we compare Semi-BABERT with several knowledge-enhanced PLMs, namely ERNIE-THU (Zhang et al., 2019), ERNIE-Baidu (Sun et al., 2019), K-BERT (Liu et al., 2020), and CKBERT (Zhang et al., 2022).

Table 7 presents the results of various PLMs on the CLUE benchmark. From an average score perspective, Semi-BABERT outperforms all PLMs, highlighting its significant advantage. Given the complexity of the CLUE benchmarks, knowledge-enhanced models typically exhibit superior performance compared to models like BERT and MacBERT. However, Semi-BABERT achieves exceptional results on this benchmark, even without additional knowledge information. Notably, Semi-BABERT surpasses the knowledge-enhanced CKBERT (70.1) and NEZHA (72.2), despite the latter models being pre-trained on larger corpus. This observation suggests that the inclusion of boundary information can compensate for the absence of extensive knowledge, demonstrating its compensatory effect on model performance.

[7] https://github.com/CLUEbenchmark/CLUE/

4. Related Work

PLMs learn sentence representations through pre-trained tasks on a large-scale corpus. For example, BERT (Devlin et al., 2019) proposes two pre-trained tasks, Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), to learn bidirectional sentence representation. Recent studies have explored extensions of the BERT model. RoBERTa (Liu et al., 2019) used strategies such as a larger corpus, a dynamic mask mechanism, and only applying MLM tasks for pre-training. ALBERT (Lan et al., 2020) reduced BERT size by sharing layers parameters and compressing word embeds. These works have achieved very successful results.

However, unlike English, Chinese lacks explicit word boundary markers such as spaces between words, posing challenges for PLM-based sequence labeling tasks such as CWS, POS and NER (Cui et al., 2021; Wei et al., 2019; Liu et al., 2021a; Jiang et al., 2022). Recent work explores methods to incorporate boundary information into Chinese PLMs. For MLM tasks, ERNIE-baidu (Sun et al., 2019) and BERT-wwm (Cui et al., 2021) apply three different masking granularities — tokens, entities, and phrases, allowing the model to learn coarse-grained boundary information at the word and phrase levels rather than just at the character level. ERNIE-Gram (Xiao et al., 2021) detects entities and phrases through statistical algorithms.

In the unsupervised approach, BABERT lever-

ages a large-scale corpus to extract a substantial amount of statistical boundary information (Jiang et al., 2022). Building upon this unsupervised boundary information, a specific unsupervised boundary-aware learning objective is designed. While these methods successfully introduce boundary information, our paper's focus lies in combining the strengths of unsupervised statistical information with supervised high-quality information. We aim to leverage the advantages of both approaches to enhance the overall performance of the model.

5. Conclusion

This paper introduces Semi-BABERT, a model specifically designed for Chinese sequence labeling tasks. Semi-BABERT incorporates lexicon-based high-quality boundary information into BABERT through a span-based boundary recognition pre-training task. Experimental results on 13 sequence labeling datasets, including tasks such as CWS, POS, and NER, demonstrate that Semi-BABERT exhibits stronger boundary awareness than other PLMs like BABERT. Furthermore, Semi-BABERT demonstrates broad effectiveness across various NLP tasks. Additionally, we propose the Boundary Information Metric (BIM), which accurately quantifies the boundary encoding potential of Chinese PLMs without task-specific fine-tuning.

6. Acknowledgement

We sincerely thank the reviewers for their invaluable feedback, which significantly improved the quality of this work. This work is supported by the National Natural Science Foundation of China (NSFC) Grant Nos.62336008 and Nos.62176180.

Jinze Bai, Shuai Bai, et al. 2023. Qwen technical report.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *CoRR*, abs/2304.00723.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. A span-extraction dataset for chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5882–5888. Association for Computational Linguistics.

Yiming Cui, Ziqing Yang, and Ting Liu. 2022. PERT: pre-training BERT with permuted language model. *CoRR*, abs/2203.06906.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2020a. ZEN: Pre-training Chinese text encoder enhanced by n-gram representations. In *Findings of the EMNLP 2020*, pages 4729–4740. Association for Computational Linguistics.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2020b. ZEN: pre-training chinese text encoder enhanced by n-gram representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4729–4740. Association for Computational Linguistics.

Ning Ding, Dingkun Long, Guangwei Xu, Muhua Zhu, Pengjun Xie, Xiaobin Wang, and Haitao Zheng. 2020. Coupling distant annotation and adversarial training for cross-domain chinese word segmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6662–6671. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language

model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13042–13054.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 601–610. ACM.

Thomas Emerson. 2005a. The second international chinese word segmentation bakeoff. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, SIGHAN@IJCNLP 2005, Jeju Island, Korea, 14-15, 2005*. ACL.

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

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wenpeng Hu, Ran Le, Bing Liu, Feng Ji, Jinwen Ma, Dongyan Zhao, and Rui Yan. 2021. Predictive adversarial learning from positive and unlabeled data. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 7806–7814. AAAI Press.

Chen Jia, Yuefeng Shi, Qinrong Yang, and Yue Zhang. 2020. Entity enhanced BERT pre-training for Chinese NER. In *Proceedings of the 2020 Conference on EMNLP*, pages 6384–6396. Association for Computational Linguistics.

Peijie Jiang, Dingkun Long, Yueheng Sun, Meishan Zhang, Guangwei Xu, and Pengjun Xie. 2021. A fine-grained domain adaption model for joint word segmentation and POS tagging. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3587–3598. Association for Computational Linguistics.

Peijie Jiang, Dingkun Long, Yanzhao Zhang, Pengjun Xie, Meishan Zhang, and Min Zhang. 2022. Unsupervised boundary-aware language model pretraining for chinese sequence labeling. *CoRR*, abs/2210.15231.

Guangjin Jin and Xiao Chen. 2008. The fourth international Chinese language processing bakeoff: Chinese word segmentation, named entity recognition and Chinese POS tagging. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

Yuxuan Lai, Yijia Liu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2021. Lattice-bert: Leveraging multi-granularity representations in chinese pre-trained language models. 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 1716–1731. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Gina-Anne Levow. 2006a. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117, Sydney, Australia. Association for Computational Linguistics.

Gina-Anne Levow. 2006b. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *Proceedings of the Fifth Workshop on Chinese Language Processing, SIGHAN@COLING/ACL 2006, Sydney, Australia, July 22-23, 2006*, pages 108–117. Association for Computational Linguistics.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Jiahuan Li, Yutong Shen, Shujian Huang, Xinyu Dai, and Jiajun Chen. 2021. When is char better than subword: A systematic study of segmentation algorithms for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 543–549, Online. Association for Computational Linguistics.

Xiaoli Li and Bing Liu. 2005. Learning from positive and unlabeled examples with different data distributions. In *Machine Learning: ECML 2005, 16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005, Proceedings*, volume 3720 of *Lecture Notes in Computer Science*, pages 218–229. Springer.

Percy Liang, Rishi Bommasani, Tony Lee, et al. 2022. Holistic evaluation of language models. *CoRR*, abs/2211.09110.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021a. Lexicon enhanced Chinese sequence labeling using BERT adapter. 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 5847–5858, Online. Association for Computational Linguistics.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021b. Lexicon enhanced chinese sequence labeling using BERT adapter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5847–5858. Association for Computational Linguistics.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: enabling language representation with knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2901–2908. AAAI Press.

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. *CoRR*, abs/1907.11692.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Ownthink. 2019. Ownthink knowledge graph.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2409–2419. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 548–554. The Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.

Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. 2021. A novel global feature-oriented relational triple extraction model based on table filling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2646–2656. Association for Computational Linguistics.

Yan Shao, Christian Hardmeier, Jörg Tiedemann, and Joakim Nivre. 2017. Character-based joint

segmentation and POS tagging for Chinese using bidirectional RNN-CRF. In *Proceedings of the 8-th IJCNLP*, pages 173–183. Asian Federation of Natural Language Processing.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180, New Orleans, Louisiana. Association for Computational Linguistics.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2020. Investigating prior knowledge for challenging chinese machine reading comprehension. *Trans. Assoc. Comput. Linguistics*, 8:141–155.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *CoRR*, abs/2304.09542.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Andrew J. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*, 13(2):260–269.

Junqiu Wei, Xiaozhe Ren, Xiaoguang Li, Wenyong Huang, Yi Liao, Yasheng Wang, Jiashu Lin, Xin Jiang, Xiao Chen, and Qun Liu. 2019. Nezha: Neural contextualized representation for chinese language understanding. *arXiv preprint arXiv:1909.00204*.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1476–1488. Association for Computational Linguistics.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Dongling Xiao, Yu-Kun Li, Han Zhang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-gram: Pre-training with explicitly n-gram masked language modeling for natural language understanding. In *Proceedings of the 2021 Conference of NAACL*, pages 1702–1715. Association for Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020a. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020b. CLUE: A chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4762–4772. International Committee on Computational Linguistics.

Nianwen Xu. 2003. Chinese word segmentation as character tagging. *Int. J. Comput. Linguistics Chin. Lang. Process.*, 8(1).

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207.

Jie Yang, Yue Zhang, and Fei Dong. 2017. Neural word segmentation with rich pretraining. 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 839–849. Association for Computational Linguistics.

Jie Yang, Yue Zhang, and Shuailong Liang. 2019a. 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, Volume*

1 (Long and Short Papers), pages 2720–2725, Minneapolis, Minnesota. Association for Computational Linguistics.

Jie Yang, Yue Zhang, and Shuailong Liang. 2019b. 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 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2720–2725. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. *CoRR*, abs/2005.07150.

Taolin Zhang, Junwei Dong, Jianing Wang, Chengyu Wang, Ang Wang, Yinghui Liu, Jun Huang, Yong Li, and Xiaofeng He. 2022. Revisiting and advancing chinese natural language understanding with accelerated heterogeneous knowledge pre-training. *CoRR*, abs/2210.05287.

Xinsong Zhang, Pengshuai Li, and Hang Li. 2021. AMBERT: A pre-trained language model with multi-grained tokenization. In *Findings of the ACL-IJCNLP 2021*, pages 421–435. Association for Computational Linguistics.

Yue Zhang and Jie Yang. 2018a. Chinese NER using lattice LSTM. In *Proceedings of the 56th ACL*, pages 1554–1564. Association for Computational Linguistics.

Yue Zhang and Jie Yang. 2018b. Chinese NER using lattice LSTM. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1554–1564. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. Chid: A large-scale chinese idiom dataset for cloze test. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 778–787. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. 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 50–61. Association for Computational Linguistics.