Unsupervised Boundary-Aware Language Model Pretraining for Chinese Sequence Labeling

Peijie Jiang[1]   **Dingkun Long**   **Yanzhao Zhang**   **Pengjun Xie**
Meishan Zhang[2*]   **Min Zhang**[2]

[1]School of New Media and Communication, Tianjin University, China
[2]Institute of Computing and Intelligence, Harbin Institute of Technology (Shenzhen)
`jzx555@tju.edu.cn`, `{zhangmeishan,zhangmin2021}@hit.edu.cn`
`{longdingkun1993,zhangyanzhao00,xpjandy}@gmail.com`

Abstract

Boundary information is critical for various Chinese language processing tasks, such as word segmentation, part-of-speech tagging, and named entity recognition. Previous studies usually resorted to the use of a high-quality external lexicon, where lexicon items can offer explicit boundary information. However, to ensure the quality of the lexicon, great human effort is always necessary, which has been generally ignored. In this work, we suggest unsupervised statistical boundary information instead, and propose an architecture to encode the information directly into pre-trained language models, resulting in Boundary-Aware BERT (BABERT). We apply BABERT for feature induction of Chinese sequence labeling tasks. Experimental results on ten benchmarks of Chinese sequence labeling demonstrate that BABERT can provide consistent improvements on all datasets. In addition, our method can complement previous supervised lexicon exploration, where further improvements can be achieved when integrated with external lexicon information.

1 Introduction

The representative sequence labeling tasks for the Chinese language, such as word segmentation, part-of-speech (POS) tagging and named entity recognition (NER) (Emerson, 2005; Jin and Chen, 2008), have been inclined to be performed at the character-level in an end-to-end manner (Shen et al., 2016). The paradigm, naturally, is standard to Chinese word segmentation (CWS), while for Chinese POS tagging and NER, it can better help reduce the error propagation (Sun and Uszkoreit, 2012; Yang et al., 2016; Liu et al., 2019a) compared with word-based counterparts by straightforward modeling.

Recently, all the above tasks have reached state-of-the-art performances with the help of BERT-like pre-trained language models (Yan et al., 2019;

Meng et al., 2019). The BERT variants, such as BERT-wwm (Cui et al., 2021), ERNIE (Sun et al., 2019), ZEN (Diao et al., 2020), NEZHA (Wei et al., 2019), etc., further improve the vanilla BERT by either using external knowledge or larger-scale training corpus. The improvements can also benefit character-level Chinese sequence labeling tasks.

Notably, since the output tags of all these character-level Chinese sequence labeling tasks involve identifying Chinese words or entities (Zhang and Yang, 2018; Yang et al., 2019), prior boundary knowledge could be highly helpful for them. A number of studies propose the integration of an external lexicon to enhance their baseline models by feature representation learning (Jia et al., 2020; Tian et al., 2020a; Liu et al., 2021). Moreover, some works suggest injecting similar resources into the pre-trained BERT weights. BERT-wwm (Cui et al., 2021) and ERNIE (Sun et al., 2019) are the representatives, which leverage an external lexicon for masked word prediction in Chinese BERT.

The lexicon-based methods have indeed achieved great success for boundary integration. However, there are two major drawbacks. First, the lexicon resources are always constructed manually (Zhang and Yang, 2018; Diao et al., 2020; Jia et al., 2020; Liu et al., 2021), which is expensive and time-consuming. The quality of the lexicon is critical to our tasks. Second, different tasks as well as different domains require different lexicons (Jia et al., 2020; Liu et al., 2021). A well-studied lexicon for word segmentation might be inappropriate for NER, and a lexicon for news NER might also be problematic for finance NER. The two drawbacks can be due to the supervised characteristic of these lexicon-based enhancements. Thus, it is more desirable to offer boundary information in an unsupervised manner.

In this paper, we propose an unsupervised Boundary-Aware BERT (BABERT), which is achieved by fully exploring the potential of statisti-

---

*Corresponding author.

cal features mined from a large-scale raw corpus. We extract a set of N-grams (a predefined fixed N) no matter they are valid words or entities, and then calculate their corresponding unsupervised statistical features, which are mostly related to boundary information. We inject the boundary information into the internal layer of a pre-trained BERT, so that our final BABERT model can approximate the boundary knowledge softly by using inside representations. The BABERT model has no difference from the original BERT, so that we can use it in the same way as the standard BERT exploration.

We conduct experiments on three Chinese sequence labeling tasks to demonstrate the effectiveness of our proposed method. Experimental results show that our approach can significantly outperform other Chinese pre-trained language models. In addition, compared with supervised lexicon-based methods, BABERT obtains competitive results on all tasks and achieves further improvements when integrated with external lexicon knowledge. We also conduct extensive analyses to understand our method comprehensively. The pre-trained model and code are publicly available at http://github.com/modelscope/adaseq/examples/babert.

Our contributions in this paper include the following: 1) We design a method to encode unsupervised statistical boundary information into boundary-aware representation, 2) propose a new pre-trained language model called BABERT as a boundary-aware extension for BERT, 3) verify BABERT on ten benchmark datasets of three Chinese sequence labeling tasks.

2 Related Work

In the past decades, machine learning has achieved good performance on sequence labeling tasks with statistical information (Bellegarda, 2004; Low et al., 2005; Bouma, 2009). Recently, neural models have led to state-of-the-art results for Chinese sequence labeling (Lample et al., 2016; Ma and Hovy, 2016; Chiu and Nichols, 2016). In addition, the presence of language representation models such as BERT (Devlin et al., 2019) has led to impressive improvements. In particular, many variants of BERT are devoted to integrating boundary information into BERT to improve Chinese sequence labeling (Diao et al., 2020; Jia et al., 2020; Liu et al., 2021).

Statistical Machine Learning Statistical information is critical for sequence labeling. Previous works attempt to count such information from large corpora in order to combine it with machine learning methods for sequence labeling (Bellegarda, 2004; Liang, 2005; Bouma, 2009). Peng et al. (2004) attempts to conduct sequence labeling by CRF and a statistical-based new word discovery method. Low et al. (2005) introduce a maximum entropy approach for sequence labeling. Liang (2005) utilizes unsupervised statistical information in Markov models, and gets a boost on Chinese NER and CWS.

Pre-trained Language Model Pre-trained language model is a hot topic in natural language processing (NLP) communities (Devlin et al., 2019; Liu et al., 2019b; Wei et al., 2019; Clark et al., 2020; Diao et al., 2020; Zhang et al., 2021) and has been extensively studied for Chinese sequence labeling. For instance, TENER (Yan et al., 2019) adopts Transformer encoder to model character-level features for Chinese NER. Glyce (Meng et al., 2019) uses BERT to capture the contextual representation combined with glyph embeddings for Chinese sequence labeling.

Lexicon-based Methods In recent studies, lexicon knowledge has been applied to improve model performance. There are two mainstream categories to the work of lexicon enhancement. The first aims to enhance the original BERT with implicit boundary information by using the multi-granularity word masking mechanism. BERT-wwm (Cui et al., 2021) and ERNIE (Sun et al., 2019) are representatives of this category, which propose to mask tokens, entities, and phrases as the mask units in the masked language modeling (MLM) task to learn the coarse-grained lexicon information during pre-training. ERNIE-Gram (Xiao et al., 2021), an extension of ERNIE, utilizes statistical boundary information for unsupervised word extraction to support masked word prediction, The second category, which includes ZEN (Diao et al., 2020), EEBERT (Jia et al., 2020), and LEBERT (Liu et al., 2021), exploits the potential of directly injecting lexicon information into BERT via extra modules, leading to better performance but is limited in predefined external knowledge. Our work follows the first line of work, most similar to ERNIE-Gram. However, different from ERNIE-Gram, we do not discretize the real-valued statistical information ex-

(a). Boundary Information Extractor (b). Boundary-Aware BERT Representation (c). Boundary-Aware BERT Learning
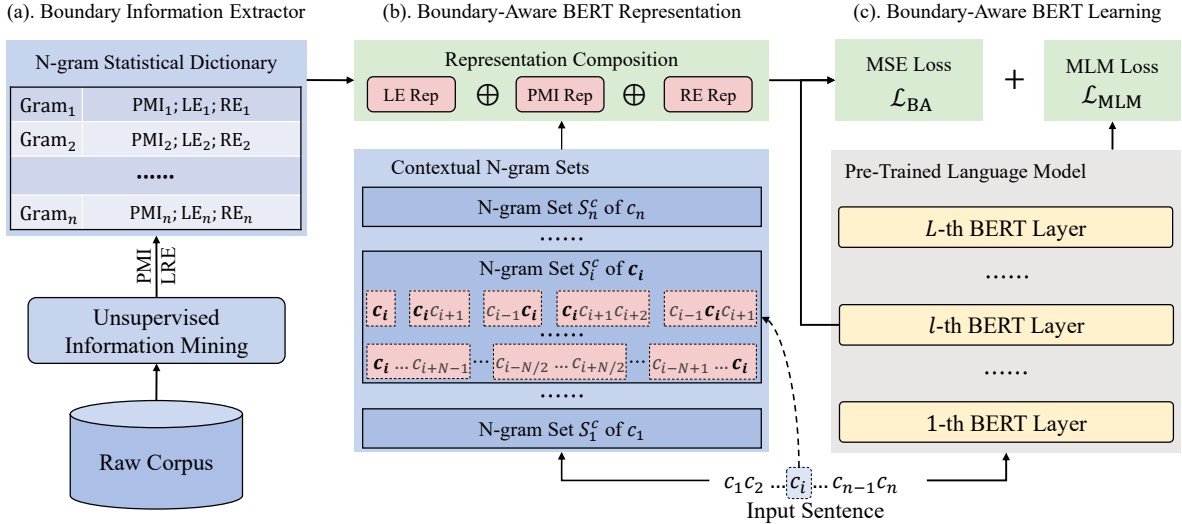
Figure 1: The overall architecture of the boundary-aware pre-trained language model, which consists of three parts: (a) boundary information extractor, (b) boundary-aware representation, and (c) boundary-aware BERT Learning. The boundary-aware objective $\mathcal{L}_{\text{BA}}$ is defined in Equation 7.

tracted from corpus, but adopt a regression manner to leverage the information fully.

3 Method

Figure 1 shows the overall architecture of our unsupervised boundary-aware pre-trained language model, which mainly consists of three components: 1) boundary information extractor for unsupervised statistical boundary information mining, 2) boundary-aware representation to integrate statistical information at the character-level, and 3) boundary-aware BERT learning which injects boundary knowledge into the internal layer of BERT. In this section, we first focus on the details of the above components, and then introduce the fine-tuning method for Chinese sequence labeling.

3.1 Boundary Information Extractor

Statistical boundary information has been shown with a positive influence on a variety of Chinese NLP tasks (Song and Xia, 2012; Higashiyama et al., 2019; Ding et al., 2020; Xiao et al., 2021). We follow this line of work, designing a boundary information extractor to mine statistical information from a large raw corpus in an unsupervised way.

The overall flow of the extractor includes two steps: I) First, we collect all N-grams from the raw corpus to build a dictionary $\mathcal{N}$, in which we count the frequencies of each N-gram and filter out the low frequencies items; II) second, considering that word frequency is insufficient for representing the flexible boundary relation in the Chinese

context, we further compute two unsupervised indicators which can capture most of the boundary information in the corpus. In the following, we will describe these two indicators in detail.

Pointwise Mutual Information (PMI) Given an N-gram, we split it into two sub-strings and compute the mutual information (MI) between them as a candidate. Then, we enumerate all sub-string pairs and choose the minimum MI as the overall PMI to estimate the tightness of the N-gram. Let $g = \{c_1...c_m\}$ be an N-gram that consists of m characters, we calculate PMI using this formula:

$$\text{PMI}(g) = \min_{i \in [1:m-1]} \{ \frac{p(g)}{p(c_1...c_i) \cdot p(c_{i+1}...c_m)} \}, \quad (1)$$

where $p(\cdot)$ denotes the probability over the corpus. Note that, when $m = 1$, the corresponding PMI is constantly equal to 1. The higher PMI indicates that the N-gram (e.g., "贝克汉姆 (Beckham)") has a similar occurrence probability to the sub-string pair (e.g., "贝克 (Beck)" and "汉姆 (Ham)"), leading to a higher association between internal sub-string pairs, which makes the N-gram more likely to be a word/entity. In contrast, a lower PMI means the N-gram (e.g., "克汉(Kehan)") is possibly an invalid word/entity.

Left and Right Entropy (LRE) Given an N-gram g, we first collect a left-adjacent character set $S_m^l = \{c_1^l, ..., c_{n_l}^l\}$ with n_l characters. Then, we utilize the conditional probability between g and its left adjacent characters in S_m^l to compute

528

the left entropy (LE), which measures sufficient boundary information. LE can be defined as:

$$\text{LE}(g) = -\sum_i^{n_l} p(c_i^l g | g) \log p(c_i^l g | g). \quad (2)$$

Similar to LE, we further collect a right adjacent set $S_m^r = \{c_1^r, ..., c_{n_r}^r\}$ with n_r characters to calculate the right entropy (RE) for the N-gram g:

$$\text{RE}(g) = -\sum_i^{n_r} p(g c_i^r | g) \log p(g c_i^r | g). \quad (3)$$

Intuitively, LRE represents the abundance of neighboring characters for the N-gram. With a lower LRE, the N-gram (e.g., "汉姆") has a more fixed context, indicating it is more likely to be a part of a phrase or entity. Conversely, the N-gram with a higher LRE (e.g., "贝克汉姆") will interact more with context, which prefers to be an independent word or phrase.

Finally, we utilize PMI and LRE to measure the flexible boundary relations in the Chinese context, and then update each N-gram in $\mathcal{N}$ with the unsupervised statistical indicators above.

3.2 Boundary-Aware Representation

By using the boundary information extractor, we can obtain an N-gram dictionary $\mathcal{N}$ with unsupervised statistical boundary information. Unfortunately, since the context independence and the high relevance to N-gram, previous works (Ding et al., 2020; Xiao et al., 2021) use such statistical features for word extraction only, which ignore the potential of statistical boundary information in representation learning. To alleviate this problem, we propose boundary-aware representation, a highly extensible method, to fully benefit from the statistical boundary information for representation learning.

To achieve boundary-aware representation, we first build contextual N-gram sets from the sentence. As shown in Figure 1 (b), given a sentence $x = \{c_1, c_2, ..., c_n\}$ with n characters and the maximum N-gram length N, we extract all N-grams that include c_i as the contextual N-gram set $S_i^c = \{\boldsymbol{c_i}, \boldsymbol{c_i} c_{i+1}, \cdots, c_{i-N+1} ... \boldsymbol{c_i}\}$ for character c_i. Then, we design a composition method to integrate the statistical features of N-grams in S_i^c by using specific conditions and rules, aiming to avoid the sparsity and contextual independence limitations of statistical information.

Concretely, we divide the information composition method into PMI and entropy representation. First, we concatenate the PMI of all N-grams in S_i^c

to generate PMI representation:

$$
\begin{aligned}
\mathbf{e}_i^p =\, & \text{PMI}(c_i) \\
& \oplus \text{PMI}(c_i c_{i+1}) \oplus \text{PMI}(c_{i-1} c_i) \\
& \oplus \text{PMI}(c_i c_{i+1} c_{i+2}) \oplus \cdots \oplus \text{PMI}(c_{i-2} c_{i-1} c_i) \quad (4) \\
& \qquad\qquad \cdots\cdots \\
& \oplus \text{PMI}(c_i ... c_{i+N-1}) \oplus \cdots \oplus \text{PMI}(c_{i-N+1} ... c_i),
\end{aligned}
$$

where $\mathbf{e}_i^p \in \mathbb{R}^a$, and $a = 1 + 2 + \cdots + N$ is the number of the N-grams that contain c_i. Note that the position of each N-gram is fixed in PMI representation. We strictly follow the order of N-gram length and the position of c_i in N-gram to concatenate their corresponding PMI, ensuring that the position and context information can be encoded into $\mathbf{e}_i^p$.

Entropy representation focuses on the contextual interactions of each character. When c_i is the border of N-grams in S_i^c, we separately aggregate the LE and RE as left and right entropy representation:

$$
\begin{aligned}
\mathbf{e}_i^{le} =\, & \text{LE}(c_i) \oplus \text{LE}(c_i c_{i+1}) \oplus \cdots \oplus \text{LE}(c_i ... c_{i+N-1}), \\
\mathbf{e}_i^{re} =\, & \text{RE}(c_i) \oplus \text{RE}(c_{i-1} c_i) \oplus \cdots \oplus \text{RE}(c_{i-N+1} ... c_i),
\end{aligned}
$$
$$(5)$$

where $\mathbf{e}_i^{le} \in \mathbb{R}^b$, $\mathbf{e}_i^{re} \in \mathbb{R}^b$, and $b = N$[1] is the number of integrated N-grams. Similar to PMI representation, the position of each N-gram in $\mathbf{e}_i^{le}$ and $\mathbf{e}_i^{le}$ is fixed and symmetric. Therefore, the boundary-aware representation $\mathbf{e}_i$ of c_i can be formalized as:

$$\mathbf{e}_i = \mathbf{e}_i^{le} \oplus \mathbf{e}_i^p \oplus \mathbf{e}_i^{re}, \quad (6)$$

where $\mathbf{e}_i \in \mathbb{R}^{a+2b}$. Finally, by composing multi-granularity statistical boundary information in a specific order, we are able to obtain the boundary-aware representation, which explicitly contains the boundary and context information.

Figure 2 shows an example of the boundary-aware representation. Given a sentence "南京市长江大桥 (Nanjing Yangtze River Bridge)" and a maximum N-gram length $N = 3$, we first build a contextual N-gram set for the character "长 (Long)". Then, we integrate the PMI of all N-grams in a specific order (from N-gram "长" to "京市长 (Mayor of Jing)") to compute PMI representation. Furthermore, left and right entropy representations are also calculated in a particular order (from N-gram "长" to "长江大 (Yangtze River Big)" and "京市长", respectively). Finally, we concatenate the above features to produce the overall boundary-aware representation of the character "长".

---

[1] We only consider the N-grams where c_i is on the boundary, which means that for each direction the number of N-grams is equal to the maximum N-gram length.
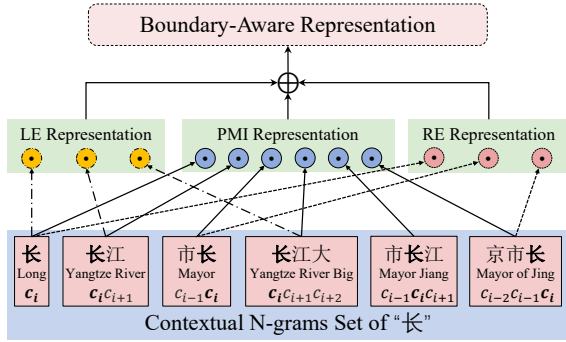
Figure 2: The illustration of boundary-aware representation for character "长" in text "南京市长江大桥".

3.3 Boundary-Aware BERT Learning

Boundary-aware BERT is a variant of BERT, enhanced with boundary information simply and effectively. In this subsection, we describe how the boundary information can be integrated into BERT during pre-training by boundary-aware learning.

Boundary-Aware Objective As mentioned in Section 3.2, given a sentence x with character-length n, we can compute the corresponding boundary-aware representation $\mathbf{E} = \{\mathbf{e_1}, ..., \mathbf{e_n}\}$. Then, we transfer the BERT feature into the boundary information space and approximate it to $\mathbf{E}$ for boundary-aware learning. Moreover, Liu et al. (2021) shows that encoding basic lexical knowledge in the shallow BERT layers is a more effective approach. Hence, we use the hidden features $\mathbf{H}^l = \{\mathbf{h}_1^l, ..., \mathbf{h}_n^l\}$ of the l-th shallow layer to achieve the boundary-aware objective:

$$\mathcal{L}_{\text{BA}} = \sum_i^n \text{MSE}(\mathbf{W}_B \mathbf{h}_i^l, \mathbf{e_i}), \qquad (7)$$

where $\text{MSE}(\cdot)$ denotes the mean square error loss. $\mathbf{W}_B$ is a trainable matrix used to project BERT representation into boundary information space.

Previous classification-based word-level masking methods use statistical information as thresholds to filter valid words for masked word prediction. Unlike the above works, we softly utilize such information in a regression manner, avoiding possible errors in empirically filtering valid tags, thereby fully exploring the potential of this information.

Pre-training Following Jia et al. (2020) and Gao and Callan (2021), we opt to initialize our model with a pre-trained BERT model released by Google[2] and randomly initialize the other parameters, alleviating the enormous cost of train-

ing BABERT from scratch. In particular, we discard the next sentence prediction task during pre-training, which is confirmed to be not essential for the pre-trained language models (Lan et al., 2020; Liu et al., 2019b). The total pre-training loss of BABERT can be formalized as:

$$\mathcal{L}_{\text{pre}} = \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{BA}}, \qquad (8)$$

where $\mathcal{L}_{\text{MLM}}$ is the standard objective of MLM task.

3.4 Fine-tuning for Sequence Labeling

Straightforward Fine-tuning As shown in Figure 1 (c), because BABERT has the same architecture as BERT, we can adopt the identical procedure that BERT uses for fine-tuning, where the output of BABERT can be used as the contextual character representation for sequence labeling. Concretely, given a sequence labeling dataset $\mathcal{D} = \{(x_j, y_j)\}_{j=1}^N$, where y_j is the label sequence of x_j, we utilize the output of BABERT and a CRF layer to calculate the sentence-level output probability $p(y_j|x_j)$, which is exactly the same as Liu et al. (2021). The negative log-likelihood loss for training can be defined as:

$$\mathcal{L}_{\text{sq}} = -\sum_j^N \log p(y_j|x_j), \qquad (9)$$

At the inference stage, we use the Viterbi algorithm (Viterbi, 1967) to generate the final label sequence.

Combining with Supervised Lexicon Features We can naturally combine BABERT with other supervised lexicon-based methods because of the unsupervised setting of BABERT. To this end, we propose a lexicon-enhanced BABERT (**BABERT-LE**) for the fine-tuning stage, which utilizes the lexicon adapter proposed by Liu et al. (2021) to incorporate external lexicon knowledge into BABERT feature:

$$\hat{\mathbf{h}}_i = \text{LA}(\mathbf{h}_i, S_i^{lex}), \qquad (10)$$

where $\text{LA}(\cdot)$ is the lexicon adapter, S_i^{lex} is a set of related N-gram embeddings of character c_i, and $\hat{\mathbf{h}}_i$ is the lexicon-enhanced version of original BERT feature $\mathbf{h}_i$. We apply the lexicon adapter after the l-th layer to be consistent with boundary-aware learning. Finally, BABERT-LE performs a similar fine-tuning procedure as BABERT for training:

$$\mathcal{L}_{\text{lex}} = -\sum_j^N \log p(y_j|x_j, [S_1^{lex}, ..., S_{n_j}^{lex}]). \qquad (11)$$

4 Experiments

4.1 Datasets

Following previous works (Devlin et al., 2019; Xiao et al., 2021), we draw the mixed corpus of Chinese Wikipedia[3] and Baidu Baike[4] as our pre-training corpus, which contains 3B tokens and 62M sentences. To further confirm the effectiveness of our proposed method for Chinese sequence labeling, we evaluate BABERT on ten benchmark datasets of three representative tasks:

Chinese Word Segmentation We use three CWS benchmarks to evaluate our BABERT. Penn Chinese TreeBank version 6.0 (CTB6) is from Xue et al. (2005), and MSRA and PKU are from SIGHAN 2005 Bakeoff (Emerson, 2005).

Part-Of-Speech Tagging For Chinese POS tagging, we conduct experiments on CTB6 (Xue et al., 2005) and the Chinese part of Universal Dependencies (UD) (Nivre et al., 2016). The UD dataset uses two different POS tagsets, which are universal and language-specific tagsets. We follow Shao et al. (2017), referring to the corpus with the two tagsets as UD1 and UD2, respectively.

Named Entity Recognition For the Chinese NER task, we conduct experiments on OntoNotes 4.0 (Onto4) (Weischedel et al., 2011) and News datasets (Jia et al., 2020), both of which are from the standard newswire domain. Moreover, we evaluate BABERT in the internet novel (Book) and financial report (Finance) domains (Jia et al., 2020) to further verify the robustness of our method.

The statistics of the benchmark datasets are shown in Table 1. For a fair comparison, we split these datasets into training, development, and test sections following previous works (Jia et al., 2020; Liu et al., 2021). Note that MSRA, PKU, and Finance do not have development sections. Therefore, we randomly select 10% instances from the training set as the development set for these datasets.

4.2 Experimental Settings

Hyperparameters During pre-training, we use the hyperparameters of $BERT_{BASE}$ to initialize BABERT and Adam (Kingma and Ba, 2014) for optimizing. The number of BERT layers L is 12, with 12 self-attention heads, 768 dimensions for hidden states, and 64 dimensions for each head.

		Train	Dev	Test
CWS	CTB6	23401	2078	2795
	MSRA	86924	-	3985
	PKU	19056	-	1944
POS	CTB6	23401	2078	2795
	UD1/2	3997	500	500
NER	Onto4	15724	4303	4346
	Book	6675	2551	2551
	News	5179	610	708
	Finance	46364	-	2000

Table 1: The statistics of sentence number for the benchmark datasets. For the datasets without development sections, we randomly select 10% sentences from the corresponding training set as the development set.

The batch size is set to 32, the learning rate is $1e$-4 with a warmup ratio of 0.1, and the max length of the input sequence is 512. To extract unsupervised boundary information, we set the maximum N-gram length N to 4 [5] and the frequency filtering threshold to 50. Then we use the 3-th BERT layer to compute boundary-aware objective. BABERT has no extra modules, which is why the parameter size and model architecture are the same as those of $BERT_{BASE}$. Finally, we train the BABERT on 8 NVIDIA Tesla V100 GPUs with 32GB memory.

For Chinese sequence labeling, we empirically set hyperparameters based on previous studies (Jia et al., 2020; Liu et al., 2021) and preliminary experiments. The batch size is 32, the max sequence length is 256, and the learning rate is fixed to $2e$-5.

Baselines To verify the effectiveness of our proposed BABERT, we build systems on the following methods to conduct fair comparisons:

- **BERT** is the Chinese version $BERT_{BASE}$ model released by Google.

- **BERT-wwm** performs segmentation on the corpus and further conduct word-level masking in pre-training (Cui et al., 2021).

- **ERNIE** is an extension of BERT, which leverages external lexicons for word-level masking (Sun et al., 2019).

- **ERNIE-Gram** is an extension of ERNIE, which alleviates the limitations of external lex-

	CWS			POS			NER				Avg
	CTB6	MSRA	PKU	CTB6	UD1	UD2	Onto4	Book	News	Finance	
(I) Pre-Trained Language Model											
BERT	97.35	98.22	96.26	94.72	95.04	94.89	80.98	76.11	79.15	85.31	89.80
BABERT	97.45	98.44	**96.70**	**95.05**	**95.65**	**95.54**	**81.90**	76.84	80.27	**86.89**	**90.47**
BERT-wwm	97.39	98.31	96.51	94.84	95.50	95.41	80.87	76.21	79.26	84.97	89.93
ERNIE	97.37	98.25	96.30	94.90	95.28	95.12	80.38	76.56	**80.36**	86.03	90.06
ERNIE-Gram	97.28	98.27	96.36	94.93	95.26	95.16	80.96	**77.19**	79.96	85.31	90.07
ZEN‡	97.33	98.28	96.55	94.76	95.55	95.54	80.06	75.67	80.17	85.05	89.90
NEZHA†	**97.53**	**98.61**	96.67	94.98	95.57	95.52	81.74	77.03	79.81	85.15	90.26
(II) Supervised Lexicon Enhanced Model											
BERT-LE$^{\ddagger\spadesuit}$	97.44	98.41	96.70	94.92	95.49	95.42	81.59	77.05	80.29	86.47	90.38
BABERT-LE‡	**97.56**	**98.63**	**96.84**	**95.24**	**95.74**	**95.70**	**82.35**	**78.36**	**80.86**	**87.25**	**90.85**
(Δ_{BABERT})	+0.11	+0.19	+0.14	+0.21	+0.09	+0.16	+0.45	+1.52	+0.59	+0.36	+0.38
Lattice‡	96.10	97.80	95.80	-	-	-	73.88	-	-	-	-
MEM-BERT‡	97.16	98.28	96.51	94.82	95.60	95.46	-	-	-	-	-
MEM-ZEN‡	97.25	98.40	96.53	94.87	95.69	95.49	-	-	-	-	-
EEBERT‡	-	-	-	-	-	-	-	77.58	76.66	87.05	-

Table 2: The overall results on three Chinese sequence labeling tasks, where we report the F1-score on the test set. $\ddagger$ denotes external knowledge is used. $\dagger$ denotes that large-scale pre-training corpus is used. $\spadesuit$ indicates that we reproduce LEBERT in a similar way for fair comparisons.

icons by using statistical information for entity and phrase extraction (Xiao et al., 2021).

- **ZEN** uses an extra N-gram encoder to integrate external lexicon knowledge into BERT during pre-training (Diao et al., 2020).

- **NEZHA** leverages functional relative positional encoding, supervised word-level masking strategy, and enormous training data[6] to enhance vanilla BERT (Wei et al., 2019).

- **BERT-LE** is a lexicon-enhanced BERT (Liu et al., 2021), which introduces a lexicon adapter between BERT layers to incorporate external lexicon embeddings. We strictly follow.Liu et al. (2021) to reimplement it with open-source word embeddings[7]

4.3 Main Results

The overall Chinese sequence labeling results are shown in Table 2. We report the F1-score of the test datasets on CWS, POS, and NER tasks. Here, we first compare our BABERT with various Chinese pre-trained language models to evaluate its effectiveness. Then, we compare BABERT-LE with other supervised lexicon-based methods to show the potential of BABERT in combining with external lexicon knowledge.

First, we examine the F1 values of the BERT baseline. As shown, BERT obtains comparable results on all Chinese sequence labeling tasks, which is similar to that of Diao et al. (2020), Tian et al. (2020a) and Liu et al. (2021). BABERT significantly outperforms BERT, resulting in an increase of $90.47 - 89.80 = 0.67$ on average. This observation clearly indicates the advantage of introducing boundary information into BERT pre-training.

Compared with various BERT extensions, our BABERT can achieve competitive performances as a whole. First, in comparison with BERT-wwm, ERNIE, ERNIE-gram, and ZEN, which leverage external lexicons that include high-frequency words for pre-training, BABERT outperforms all of them by averaging $\frac{0.54+0.41+0.40+0.57}{4} = 0.48$ point, and achieves top scores on eight of the ten benchmarks. This result is consistent with our intuition that directly exploiting a supervised lexicon can only achieve good performance in specific tasks, indicating the limitation of these methods when the chosen lexicon is incompatible with the target tasks. Second, we find that BABERT surpasses NEZHA in the average F1 values, indicating that the boundary information is more critical than the data scale for Chinese sequence labeling.

Then, we compare our method with supervised lexicon-based methods. Lattice-based methods (Zhang and Yang, 2018; Yang et al., 2019) are the first to integrate word features into neural networks for Chinese sequence labeling. MEM (Tian et al.,

---

[6]NEZHA uses three large corpora, including Chinese Wikipedia, Baidu Baike, and Chinese News, which contain 11B tokens and are four times more than us.

[7]https://ai.tencent.com/ailab/nlp/en/embedding.html

	CWS-PKU			NER-Onto4		
	10	50	100	10	50	100
BERT	83.97	87.93	88.25	14.87	42.37	57.95
BABERT	**84.73**	**89.45**	**90.02**	**32.07**	**46.55**	**60.58**
BERT-wwm	84.70	88.05	88.84	12.75	43.09	59.44
ERNIE	84.31	87.04	88.20	19.86	42.96	50.81
ERNIE-Gram	84.01	86.62	87.99	28.39	45.88	60.01
NEZHA	84.40	88.70	89.73	14.45	44.10	59.20

Table 3: Few-shot results on PKU and Onto4, using 10, 50, and 100 instances of the training data.

2020a,b) designs an external memory network after the BERT encoder to incorporate lexicon knowledge. EEBERT (Jia et al., 2020) builds entity embeddings from the corpus and further utilizes them in the multi-head attention mechanism. The results are shown in Table 2 (II). All the above methods lead to significant improvements over the base BERT model, which shows the effectiveness of external lexicon knowledge. Moreover, BABERT can achieve comparable performance with the above methods, which further demonstrates the potential of our unsupervised manner.

BABERT learns boundary information from unsupervised statistical features with vanilla BERT, which means it has excellent scalability to fuse with other BERT-based supervised lexicon models. As shown, we can see that our BABERT-LE achieves further improvements and state-of-the-art performances on all tasks, showing the advantages of our unsupervised setting and boundary-aware learning. Interestingly, compared with MEM-ZEN, BABERT-LE has larger improvements over their corresponding baselines. One reason might be that both ZEN and the memory network module exploits supervised lexicons, which leads to a duplication of introduced knowledge.

4.4 Analysis

In this subsection, we conduct detailed experimental analyses for an in-depth comprehensive understanding of our method.

Few-Shot Setting To further verify the effectiveness of BABERT, we conduct experiments under the few-shot setting, where we randomly sample 10, 50, and 100 instances of the original training data from PKU (CWS) and Onto4 (NER). For fair comparisons, we compare BABERT with the pretrained language models without external supervised knowledge. The results are presented in Table 3. As the size of training data is reduced, the

	Onto4	Book	News	Finance
BABERT	**81.90**	**76.84**	**80.27**	**86.89**
+ T-test	81.77	76.65	80.12	86.23
- PMI	81.12	76.28	79.51	85.56
- LRE	81.51	76.47	79.67	85.94
- LE	81.58	76.62	79.88	85.94
- RE	81.53	76.59	79.75	85.90

Table 4: The results of different feature combining settings on the four NER benchmark datasets.

Layer	CWS PKU	POS CTB6	NER Onto4	NER News
12	96.38	94.68	81.02	79.19
6	96.53	94.84	81.45	79.91
3	**96.70**	**95.03**	**81.90**	**80.27**
1	96.62	94.90	81.59	79.52

Table 5: The influence of boundary information learn at different layers of BERT model.

performance drops significantly, indicating that the performances of such models rely on the scale of the labeled training data. Nevertheless, BABERT achieves top scores under each setting and significantly outperforms vanilla BERT, demonstrating the potential of injecting unsupervised boundary information by using regression-based boundary-aware learning, which effectively alleviates the low-resource problem.

Choice of Statistical Features As mentioned in Section 3.2, we use PMI and LRE to model unsupervised boundary-aware representation. In addition to these features, T-test (Xiao et al., 2021) is another popular choice that can be utilized. Thus, we conduct ablation experiments to check the effectiveness of these features. We analyze five feature combining settings, including PMI (-LRE) and LRE (-PMI) alone, the ablation study of LRE (-LE and -RE), and our two indicators with additional T-test features being concatenated at the end of the original boundary-aware representation (+T-test). As shown in Table 4, the combination of PMI and LRE can achieve the best results, discarding either of them will result in decreased performance. Besides, right entropy is more important than left entropy according to our results, which may be that the right entropy is more compatible with the reading characteristics of Chinese. Interestingly,

	ENT	GAM	LOT	FIN
BERT	89.50	74.35	83.62	79.78
BABERT	**89.86**	**75.65**	83.74	80.22
BERT-wwm	83.82	75.39	81.18	**81.38**
ERNIE	89.75	74.76	86.69	80.20
ERNIE-Gram	88.95	74.58	84.21	79.83
NEZHA	89.47	72.78	**87.12**	79.75

Table 6: The results of different feature combining settings on the four NER benchmark datasets.

Sentence	如果 WCG2011$_{org}$有 WAR3$_{game}$和 星际2$_{game}$两个项目 If WCG2021 has WAR3 and StarCraft2 these two projects
BERT	如果 WCG$_{org}$2011有 WAR3$_{game}$和 星际2$_{game}$两个项目
BABERT	如果 WCG2011$_{org}$有 WAR3$_{game}$和 星际2$_{game}$两个项目
BERT-wwm	如果 WCG2$_{org}$011有 WAR3$_{game}$和 星际2$_{game}$两个项目
ERNIE	如果 WCG20$_{org}$11有 WAR3$_{game}$和 星际2$_{game}$两个项目
ERNIE-Gram	如果 WCG$_{org}$2011有 WAR3$_{game}$和 星际2$_{game}$两个项目
NEZHA	如果WCG2011有 WAR3$_{org}$和 星际2$_{game}$两个项目

Table 7: Example from the game domain. Red (Blue) represents correct (incorrect) entities.

adding T-test does not bring further improvements. One possible reason is that the T-test is essentially similar to the entropy measure of 2-grams, which has already been injected into our BABERT model.

Boundary Information Encoding Layer Previous works (Jawahar et al., 2019; Liu et al., 2021) exploit the fact that different BERT layers would generate different concept representations. The shallow BERT layers are more likely to capture basic lexicon information, while the top layers focus on the semantic representation. We empirically set l in $\{1, 3, 6, 12\}$ to explore the effect of computing boundary-aware loss by the hidden features $\mathbf{H}^l$ of different BERT layers on Chinese sequence labeling tasks. Table 5 shows the results. We can see that the best F1-score can be achieved when $l = 3$ on all datasets, which indicates that the BABERT still needs sufficient parameters to learn the basic boundary information. Interestingly, the BABERT performs poorly when $l = 12$, which might be due to a conflict between the MLM loss and our boundary-aware regression loss during pretraining.

Qualitative Analysis To explore how BABERT improves the performance for Chinese sequence labeling, we conduct qualitative analysis on the News test dataset, which consists of four different subdomains, namely game (GAM), entertainment (ENT), lottery (LOT) and finance (FIN). The results are shown in Table 6. We can see that compared with other pre-trained language models, BABERT can obtain consistent improvement in all domains with unsupervised statistical boundary information, while the other models only improve performance on specific domains. Moreover, as shown in Table 7, we also give an example from the game domain to further demonstrate the effectiveness of our method. BABERT is the only model that correctly recognizes all entities. In particular, the prediction of BABERT for the entity "WCG2011$_{org}$" indicates the potential of boundary information.

5 Conclusion

In this paper, we proposed BABERT, a novel unsupervised boundary-aware pre-training model for Chinese sequence labeling. In BABERT, given a Chinese sentence, we calculated boundary-aware representation with unsupervised statistical information to capture boundary information, and directly injected such information into BERT weights during pre-training. Unlike previous works, BABERT exploited an effective way to utilize boundary information in an unsupervised manner, thereby alleviating the limitations of supervised lexicon-based approaches. Experimental results on ten benchmark datasets of three different tasks illustrated that our method was highly effective and better than other Chinese pre-trained models. Moreover, the combination with supervised lexicon extensions could achieve further improvements and state-of-the-art results on most tasks.

Limitations

BABERT suffers from three major limitations. The first limitation is that in the boundary information extractor, we empirically chose PMI and LRE. In addition to these indicators and the T-test measure we verified in the experimental analyses, some alternatives that contain boundary information could be used to compute boundary-aware representation. Thus, we plan to explore more unsupervised statistical features. The second limitation is that we focused only on Chinese sequence labeling tasks in this work, ignoring the potential of boundary information and BABERT in other Chinese NLP tasks. The third one is that we only consider BABERT for Chinese. For other languages which do not use spaces between words such as Japanese and Thai, we can also attempt to inject boundary information, and the effectiveness in these languages should be verified by experiments. Further research is needed to evaluate our BABERT in future studies.

Acknowledgements

References

Jerome R Bellegarda. 2004. Statistical language model adaptation: review and perspectives. *Speech communication*, 42(1):93–108.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

Jason Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4(0):357–370.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of 2019 Conference of the NAACL*, pages 4171–4186.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2020. ZEN: Pre-training Chinese text encoder enhanced by n-gram representations. In *Findings of the EMNLP 2020*, 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 ACL*, pages 6662–6671. 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*.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 981–993. Association for Computational Linguistics.

Shohei Higashiyama, Masao Utiyama, Eiichiro Sumita, Masao Ideuchi, Yoshiaki Oida, Yohei Sakamoto, and Isaac Okada. 2019. Incorporating word attention into character-based word segmentation. In *Proceedings of the 2019 Conference of the NAACL*, pages 2699–2709. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th ACL*.

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.

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*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of NAACL*, pages 260–270. 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 *International Conference on Learning Representations*.

Percy Liang. 2005. *Semi-supervised learning for natural language*. Ph.D. thesis, Massachusetts Institute of Technology.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021. Lexicon enhanced Chinese sequence labeling using BERT adapter. In *Proceedings of the 59th ACL and the 11th IJCNLP*, pages 5847–5858. Association for Computational Linguistics.

Wei Liu, Tongge Xu, Qinghua Xu, Jiayu Song, and Yueran Zu. 2019a. An encoding strategy based word-character LSTM for Chinese NER. In *Proceedings of the 2019 Conference of the NAACL*, pages 2379–2389. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jin Kiat Low, Hwee Tou Ng, and Wenyuan Guo. 2005. A maximum entropy approach to chinese word segmentation. In *Proceedings of the fourth SIGHAN workshop on Chinese language processing*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th ACL*, pages 1064–1074. Association for Computational Linguistics.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. 2019. Glyce: Glyph-vectors for chinese character representations. *Advances in Neural Information Processing Systems*, 32.

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.

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

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.

Mo Shen, Wingmui Li, HyunJeong Choe, Chenhui Chu, Daisuke Kawahara, and Sadao Kurohashi. 2016. Consistent word segmentation, part-of-speech tagging and dependency labelling annotation for Chinese language. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 298–308. The COLING 2016 Organizing Committee.

Yan Song and Fei Xia. 2012. Using a goodness measurement for domain adaptation: A case study on chinese word segmentation. In *In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC' 12*.

Weiwei Sun and Hans Uszkoreit. 2012. Capturing paradigmatic and syntagmatic lexical relations: Towards accurate Chinese part-of-speech tagging. In *Proceedings of the 50th ACL*, pages 242–252. Association for Computational Linguistics.

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*.

Yuanhe Tian, Yan Song, Xiang Ao, Fei Xia, Xiaojun Quan, Tong Zhang, and Yonggang Wang. 2020a. Joint Chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge. In *Proceedings of the 58th ACL*, pages 8286–8296. 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 ACL*, pages 8274–8285. Association for Computational Linguistics.

Andrew Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information 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*.

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.

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.

Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. 2019. Tener: adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*.

Jie Yang, Zhiyang Teng, Meishan Zhang, and Yue Zhang. 2016. Combining discrete and neural features for sequence labeling. In *The 17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*.

Jie Yang, Yue Zhang, and Shuailong Liang. 2019. Subword encoding in lattice LSTM for Chinese word segmentation. In *Proceedings of the 2019 Conference of the NAACL*, pages 2720–2725. Association for Computational Linguistics.

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