Enhancing Pre-trained Chinese Character Representation with Word-aligned Attention

Yanzeng Li, Bowen Yu, Mengge Xue, Tingwen Liu*
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
{liyanzeng, yubowen, xuemengge, liutingwen}@iie.ac.cn

Abstract

Most Chinese pre-trained models take character as the basic unit and learn representation according to character’s external contexts, ignoring the semantics expressed in the word, which is the smallest meaningful utterance in Chinese. Hence, we propose a novel word-aligned attention to exploit explicit word information, which is complementary to various character-based Chinese pre-trained language models. Specifically, we devise a pooling mechanism to align the character-level attention to the word level and propose to alleviate the potential issue of segmentation error propagation by multi-source information fusion. As a result, word and character information are explicitly integrated at the fine-tuning procedure. Experimental results on five Chinese NLP benchmark tasks demonstrate that our method achieves significant improvements against BERT, ERNIE and BERT-wwm.

1 Introduction

Pre-trained language Models (PLM) such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), ERNIE (Sun et al., 2019), BERT-wwm (Cui et al., 2019) and XLNet (Yang et al., 2019) have been proven to capture rich language information from text and then benefit many NLP applications by simple fine-tuning, including sentiment classification (Pang et al., 2002), natural language inference (Bowman et al., 2015), named entity recognition (Sang and De Meulder, 2003) and so on.

Generally, most popular PLMs prefer to use attention mechanism (Vaswani et al., 2017) to represent the natural language, such as word-to-word self-attention for English. Unlike English, in Chinese, words are not separated by explicit delimiters. Since without word boundaries information, it is intuitive to model characters in Chinese tasks directly. However, in most cases, the semantic of a single Chinese character is ambiguous. For example, the character “拍” in word “球拍 (bat)” and “拍卖 (auction)” has entirely different meanings. Moreover, several recent works have demonstrated that considering the word segmentation information can lead to better language understanding, and accordingly benefits various Chinese tasks (Wang et al., 2017; Li et al., 2018; Zhang and Yang, 2018; Gui et al., 2019; Mengge et al., 2019).

All these factors motivate us to expand the character-level attention mechanism in Chinese PLMs to represent the semantics of words. To this end, there are two main challenges. (1) How to seamlessly integrate the segmentation information into character-based attention module of PLM is an important problem. (2) Gold-standard segmentation is rarely available in the downstream tasks, and how to effectively reduce the cascading noise caused by Chinese word segmentation (CWS) tools (Li et al., 2019) is another challenge.

In this paper, we propose a new architecture, named Multi-source Word Aligned Attention (MWA), to solve the above issues. (1) Psycholinguistic experiments (Bai et al., 2008; Meng et al., 2014) have shown that readers are likely to pay approximate attention to each character in one Chinese word. Drawing inspiration from such findings, we introduce a novel word-aligned attention, which could aggregate attention weight of characters in one word into a unified value with the mixed pooling strategy (Yu et al., 2014). (2) For reducing segmentation error, we further extend our word-aligned attention with multi-source segmentation produced by various segmenters and deploy

---

*Corresponding author
The primary goal of this work is to inject the word aligned layer on top of the encoder to integrate the text, it neglects the semantic information expressed able ability to capture language knowledge from enriched contextual representations. In this work, and the outputs from the last dated variants (ERNIE, BERT-wwm) as the basic on language modeling, we adopt BERT and its up-

strong performance of deep Transformers trained nese PLMs and enhance original models. Given the segmentation knowledge into character-level Chi-

2.1 Character-level Pre-trained Encoder

The primary goal of this work is to inject the word segmentation knowledge into character-level Chi-

nese PLMs and enhance original models. Given the strong performance of deep Transformers trained on language modeling, we adopt BERT and its up-

dated variants (ERNIE, BERT-wwm) as the basic encoder in this work, and the outputs from the last layer of encoder are treated as the character-level enriched contextual representations $H$.

2.2 Word-aligned Attention

Although character-level Chinese PLM has remarkable ability to capture language knowledge from text, it neglects the semantic information expressed in the word level. Therefore we apply a word-aligned layer on top of the encoder to integrate the word boundary information into the representation of characters with an attention aggregation module.

For an input sequence with $n$ characters $S = [c_1, c_2, ..., c_n]$, where $c_j$ denotes the $j$-th character, CWS tool $\pi$ is used to partition $S$ into non-overlapping word blocks: $\pi(S) = [w_1, w_2, ..., w_m], (m \leq n)$ (1)

where $w_i = \{c_s, c_{s+1}, ..., c_{s+l-1}\}$ is the $i$-th segmented word with a length of $l$ and $s$ is the index of $w_i$’s first character in $S$. We apply self-attention operation with the representations of all input characters to get the character-level attention score matrix $A_c \in \mathbb{R}^{n \times n}$. It can be formulated as:

$$A_c = \mathcal{F}(H) = \text{softmax}\left(\frac{(KW_k)(QW_q)^T}{\sqrt{d}}\right)$$ (2)

where $Q$ and $K$ are both equal to the collective representation $H$ at the last layer of the Chinese PLM, $W_k \in \mathbb{R}^{d \times d}$ and $W_q \in \mathbb{R}^{d \times d}$ are trainable parameters for projection. While $A_c$ models the relationship between two arbitrarily characters without regard to the word boundary, we argue that incorporating word as atom in the attention can better represent the semantics, as the literal meaning of each individual character can be quite different from the implied meaning of the whole word, and the simple weighted sum in the character level may lose word and word sequence information.

To address this issue, we propose to align $A_c$ in the word level and integrate the inner-word attention. For ease of exposition, we rewrite $A_c$ as $[a_{c,1}^1, a_{c,2}^1, ..., a_{c,n}^1]$, where $a_{c,i}^k \in \mathbb{R}^n$ denotes the $i$-th row vector of $A_c$, that is, $a_{c,i}^k$ represents the attention score vector of the $i$-th character. Then we deploy $\pi$ to segment $A_c$ according to $\pi(S)$. For example, if $\pi(S) = [\{c_1, c_2\}, \{c_3\}, ..., \{c_{n-1}, c_n\}]$, then

$$\pi(A_c) = [\{a_{c,1}^1, a_{c,2}^1\}, a_{c,3}^1, ..., \{a_{c,n-1}^{n-1}, a_{c,n}^n\}]$$ (3)

In this way, an attention vector sequence is divided into several subsequences and each subsequence represents the attention of one word.

<table>
<thead>
<tr>
<th>Chinese</th>
<th>北京西山森林公园</th>
</tr>
</thead>
<tbody>
<tr>
<td>thulac</td>
<td>北京 西山 森林 公园</td>
</tr>
<tr>
<td>ictclas</td>
<td>北京 西山 森林 公园</td>
</tr>
<tr>
<td>hanlp</td>
<td>北京 西山 森林 公园</td>
</tr>
</tbody>
</table>

Table 1: Results of different popular CWS tools over “北京西山森林公园 (Beijing west mount forest park)”.

Figure 1: Architecture of Word-aligned Attention

The source code of this paper can be obtained from https://github.com/lsvih/MWA.
Then, motivated by the psycholinguistic finding that readers are likely to pay similar attention to each character in one Chinese word, we devise an appropriate aggregation module to fuse the inner-word attention. Concretely, we first transform \( \{a_{c}^{s}, ..., a_{c}^{s+l-1}\} \) into one attention vector \( a_{w}^{i} \) for \( w_{i} \) with the mixed pooling strategy (Yu et al., 2014). Then we execute the piecewise upsampling operation over each \( a_{w}^{i} \) to keep input and output dimensions unchanged for the sake of plug and play. The detailed process can be summarized as:

\[
\hat{a}_{w}^{i} = \lambda \text{Maxpooling}(\{a_{c}^{s}, ..., a_{c}^{s+l-1}\}) + (1 - \lambda) \text{Meanpooling}(\{a_{c}^{s}, ..., a_{c}^{s+l-1}\})
\]

\[
\hat{A}_{c}[s : s + l - 1] = e_{l} \otimes a_{w}^{i}
\]

where \( \lambda \in R^{1} \) is a weighting trainable variable to balance the mean and max pooling, \( e_{l} = [1, ..., 1]^{T} \) represents a \( l \)-dimensional all-ones vector, \( l \) is the length of \( w_{i} \), \( e_{l} \otimes a_{w}^{i} = [a_{w}^{i}, ..., a_{w}^{i}] \) denotes the kronecker product operation between \( e_{l} \) and \( a_{w}^{i} \), \( \hat{A}_{c} \in R^{n \times n} \) is the aligned attention matrix. Eqs. 4 and 5 can help incorporate word segmentation information into character-level attention calculation process, and determine the attention vector of one character from the perspective of the whole word, which is beneficial for eliminating the attention bias caused by character ambiguity. Finally, we can obtain the enhanced character representation produced by word-aligned attention as follows:

\[
\hat{H} = \hat{A}_{c}VW_{v}
\]

where \( V = H, W_{v} \in R^{dx} \) is a trainable projection matrix. Besides, we also use multi-head attention (Vaswani et al., 2017) to capture information from different representation subspaces jointly, thus we have \( K \) different aligned attention matrices \( \hat{A}_{c}^{k} (1 \leq k \leq K) \) and corresponding representation \( H^{k} \). With multi-head attention architecture, the output can be expressed as follows:

\[
H = \text{Concat}(H^{1}, H^{2}, ..., H^{K})W_{o}
\]

### 2.3 Multi-source Word-aligned Attention

As mentioned in Section 1, our proposed word-aligned attention relies on the segmentation results of CWS tool \( \pi \). Unfortunately, a segmenter is usually unreliable due to the risk of ambiguous and non-formal input, especially on out-of-domain data, which may lead to error propagation and an unsatisfactory model performance. In practice, the ambiguous distinction between morphemes and compound words leads to the cognitive divergence of words concepts, thus different \( \pi \) may provide diverse \( \pi(S) \) with various granularities. To reduce the impact of segmentation error and effectively mine the common knowledge of different segmenters, it’s natural to enhance the word-aligned attention layer with multi-source segmentation inputs. Formally, assume that there are \( M \) popular CWS tools employed, we can obtain \( M \) different representations \( H^{1}, ..., H^{M} \) by Eq. 7. Then we propose to fuse these semantically different representations as follows:

\[
\bar{H} = \sum_{m=1}^{M} \tanh(H^{m}W_{g})
\]

where \( W_{g} \) is a parameter matrix and \( \bar{H} \) denotes the final output of the MWA attention layer.

### 3 Experiments

#### 3.1 Experiments Setup

To test the applicability of the proposed MWA attention, we choose three publicly available Chinese pre-trained models as the basic encoder: BERT, ERNIE, and BERT-wwm. In order to make a fair comparison, we keep the same hyper-parameters (such maximum length, warm-up steps, initial learning rate, etc.) as suggested in BERT-wwm (Cui et al., 2019) for both baselines and our method on each dataset. We run the same experiment for five times and report the average score to ensure the reliability of results. Besides, three popular CWS tools: thulac (Sun et al., 2016), ictclas (Zhang et al., 2003) and hanlp (He, 2014) are employed to segment sequence.

The experiments are carried out on five Chinese NLP tasks and six public benchmark datasets:

- **Sentiment Classification (SC)**: We adopt ChnSentiCorp\(^{4}\) and weibo-100k sentiment dataset\(^{5}\) in this task. ChnSentiCorp dataset has about 10k sentences, which express positive or negative emotion. weibo-100k dataset contains 1.2M microblog

\(^{4}\)https://github.com/pengming617/bert_classification

\(^{5}\)https://github.com/SophonPlus/ChineseNlpCorpus/
texts and each microblog is tagged as positive or negative emotion.

**Named Entity Recognition (NER):** this task is to test model’s capacity of sequence tagging. We use a common public dataset Ontonotes 4.0 (Weischedel et al., 2011) in this task.

**Sentence Pair Matching (SPM):** We use the most widely used dataset LCQMC (Liu et al., 2018) in this task, which aims to identify whether two questions are in a same intention.

**Natural Language Inference (NLI):** this task is to exploit the contexts of text and concern inference relationships between sentences. XNLI (Conneau et al., 2018) is a cross-language language understanding dataset; we only use the Chinese language part of XNLI to evaluate the language understanding ability. And we processed this dataset in the same way as ERNIE (Sun et al., 2019) did.

**Machine Reading Comprehension (MRC):** MRC is a representative document-level modeling task which requires to answer the questions based on the given passages. DRCD (Shao et al., 2018) is a public span-extraction Chinese MRC dataset, whose answers are spans in the document.

We implement our model with PyTorch (Paszke et al., 2019), and all baselines are converted weights into PyTorch version. All experiments employ modified Adam (Devlin et al., 2019) as optimizer with 0.01 weight decay and 0.1 warm-up ratio. All pre-trained models are configured to 12 layers and 768 hidden dimension. The detail settings are shown in Table 2.

### 3.2 Experiment Results

Table 3 shows the performances on five classical Chinese NLP tasks with six public datasets. Generally, our method consistently outperforms all baselines on all five tasks, which demonstrates the effectiveness and universality of the proposed approach. Moreover, the Wilcoxon’s test shows that a significant difference ($p < 0.05$) exits between our model and baseline models.

In detail, on the two datasets of SC task, we observe an average of 0.53% and 0.83% absolute improvement in F1 score, respectively. SPM and NLI tasks can also gain benefits from our enhanced representation. For the NER task, our method obtains 0.92% improvement averagely over all baselines. Besides, introducing word segmentation information into the encoding of character sequences improves the MRC performance on average by 1.22 points and 1.65 points in F1 and Exact Match (EM) score respectively. We attribute such significant gain in NER and MRC to the particularity of these two tasks. Intuitively, Chinese NER is correlated with word segmentation, and named entity boundaries are also word boundaries. Thus the potential boundary information presented by the additional segmentation input can provide better guidance to label each character, which is consistent with the conclusion in (Zhang and Yang, 2018). Similarly, the span-extraction MRC task is to extract answer spans from document (Shao et al., 2018), which also faces the same word boundary problem as NER, and the long sequence in MRC exacerbates the problem. Therefore, our method gets a relatively greater improvement on the DRCD dataset.

### 3.3 Ablation Study

To demonstrate the effectiveness of our multi-source fusion method, we carry out experiments on the DRCD dev set with different segmentation inputs. Besides, we also design two strong baselines by introducing a Transformer layer ($1T$) and a random tokenizer model ($WArandom$) to exclude the benefits from additional parameters. As shown in Table 4, adding additional parameters by introducing an extra transformer layer can benefit the PLMs. Compared with $1T$ and $WArandom$, our proposed word-aligned attention gives quite stable improvements no matter what CWS tool we use, which again confirms the effectiveness and rationality of incorporating word segmentation information into character-level PLMs. Another observation is that
Table 3: Evaluation results regarding each model on different datasets. Bold marks highest number among all models. Numbers in brackets indicate the absolute increase over baseline models. Superscript number 1, 2, 3, 4 respectively represents that the corresponding dataset is also used by BERT (Devlin et al., 2019), BERT-wwm (Wu et al., 2016; Cui et al., 2019), ERNIE (Sun et al., 2019) and Glyce (Meng et al., 2019a), respectively. The results of all baselines are produced by our implementation or retrieved from original papers, and we report the higher one among them. The improvements over baselines are statistically significant ($p < 0.05$). † denotes the results of previous state-of-the-art models on these datasets without using BERT.

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
<th>Param. Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-wwm</td>
<td>91.68</td>
<td>110M</td>
</tr>
<tr>
<td>BERT-wwm+1T</td>
<td>92.23</td>
<td>110M+7.1M</td>
</tr>
<tr>
<td>BERT-wwm+2T</td>
<td>91.99</td>
<td>110M+14.2M</td>
</tr>
<tr>
<td>BERT-wwm+3T</td>
<td>91.68</td>
<td>110M+21.3M</td>
</tr>
<tr>
<td>BERT-wwm+MWA</td>
<td>93.72</td>
<td>110M+7.6M</td>
</tr>
<tr>
<td>Robust-BERT-wwm-ext-large</td>
<td>94.40</td>
<td>340M</td>
</tr>
</tbody>
</table>

Table 4: F1 results of ablation experiments on the DRCD dev set.

employing multiple segmenters and fusing them together could introduce richer segmentation information and further improve the performance.

### 3.4 Parameter Scale Analysis

For fair comparison and demonstrating the improvement of our model is not only rely on more trainable parameters, we also conduct experiments on the DRCD dev set to explore whether the performance keeps going-up with more parameters by introducing additional transformer blocks on top of the representations of PLMs.

In Table 5, +1T denotes that we introduce another one Transformer layer on top of BERT-wwm and +2T means additional 2 layers, $M$ denotes million. As the experimental results showed, when the number of additional layers exceeds 1, the performance starts to decline, which demonstrates that using an extensive model on top of the PLM representations may not bring additional benefits. We can conclude that MWA doesn’t introduce too many parameters, and MWA achieves better performance than +1T under the similar parameter numbers. Besides, we also make comparison with the current best Chinese PLM: Robust-BERT-wwm-ext-large (Cui et al., 2019), a 24-layers Chinese PLM with 13.5 times more pre-training data and 3.1 times more parameters than BERT-wwm, experimental results show that our model can achieve comparable performance, which again confirms the effectiveness of incorporating word segmentation information into character-level PLMs.

### 4 Conclusion

In this paper, we develop a novel Multi-source Word Aligned Attention model (referred as MWA), which integrates word segmentation information into character-level self-attention mechanism to enhance the fine-tuning performance of Chinese PLMs. We conduct extensive experiments on five NLP tasks with six public datasets. The proposed approach yields substantial improvements compared to BERT, BERT-wwm and ERNIE, demonstrating its effectiveness and universality. Furthermore, the word-aligned attention can also be applied to English PLMs to bridge the semantic gap between the whole word and the segmented WordPiece tokens, which we leave for future work.

### Acknowledgement

We would like to thank reviewers for their insightful comments. This work is supported by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDC02040400.
References


Han He. 2014. HanLP: Han Language Processing.


Xiaoya Li, Yuxian Meng, Xiaofei Sun, Qinghong Han, Arianna Yuan, and Jiwei Li. 2019. Is word segmentation necessary for deep learning of chinese representations? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3242–3252.


