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

Anonymous ACL submission

Abstract

Most Chinese pre-trained models take character as 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 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 directly model characters in Chinese tasks. However, in most cases, the semantic of 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).

All these factors motivate us to expand the character-level attention mechanism in Chinese PLMs to represent the semantics of words [1]. 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 **M**ulti-source **W**ord **A**ligned 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 finding, 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 a fusion function to pull together their disparate outputs. As shown in Table 1, different CWS tools

[1] Considering the enormous cost of re-training a language model, we hope to incorporate word segmentation information to the fine-tuning process to enhance performance, and leave how to improve the pre-training procedure for a future work.

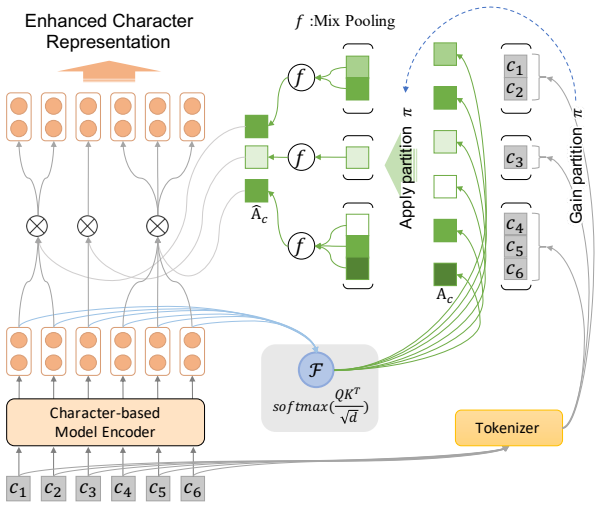

Figure 1: Architecture of Word-aligned Attention

may have different annotation granularity. Through comprehensive consideration of multi-granularity segmentation results, we can implicitly reduce the error caused by automatic annotation.

Extensive experiments are conducted on various Chinese NLP tasks including named entity recognition, sentiment classification, sentence pair matching, natural language inference and machine reading comprehension. The results and analysis show that the proposed method boosts BERT, ERNIE and BERT-wwm significantly on all the datasets[2].

2 Methodology

2.1 Character-level Pre-trained Encoder

The primary goal of this work is to inject the word segmentation knowledge into character-level Chinese PLMs and enhance original models. Given the strong performance of deep Transformers trained on language modeling, we adopt BERT and its updated 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.

[2]The source code of this paper can be obtained from https://github.com/xxx/xxx.

Chinese		北京西山森林公园				
Segmenter	thulac	北京	西山		森林	公园
	ictclas	北京	西	山	森林	公园
	hanlp	北京	西山		森林公园	

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

For an input sequence with n characters $S = [c_1, c_2, ..., c_n]$, where c_j denotes the j-th character, CWS tool π is used to partition S into non-overlapping word blocks:

$$\pi(S) = [w_1, w_2, ..., w_m], (m \leq n) \quad (1)$$

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 $\mathbf{A}_c \in \mathbb{R}^{n \times n}$. It can be formulated as:

$$\mathbf{A}_c = \mathcal{F}(\mathbf{H}) = \text{softmax}(\frac{(\mathbf{K}\mathbf{W}_k)(\mathbf{Q}\mathbf{W}_q)^T}{\sqrt{d}}), \quad (2)$$

where $\mathbf{Q}$ and $\mathbf{K}$ are both equal to the collective representation $\mathbf{H}$ at the last layer of the Chinese PLM, $\mathbf{W}_k \in \mathbb{R}^{d \times d}$ and $\mathbf{W}_q \in \mathbb{R}^{d \times d}$ are trainable parameters for projection. While $\mathbf{A}_c$ 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 $\mathbf{A}_c$ in the word level and integrate the inner-word attention. For ease of exposition, we rewrite $\mathbf{A}_c$ as $[\mathbf{a}_c^1, \mathbf{a}_c^2, ..., \mathbf{a}_c^n]$, where $\mathbf{a}_c^i \in \mathbb{R}^n$ denotes the i-th row vector of $\mathbf{A}_c$, that is, $\mathbf{a}_c^i$ represents the attention score vector of the i-th character. Then we deploy π to segment $\mathbf{A}_c$ according to $\pi(S)$. For example, if $\pi(S) = [\{c_1, c_2\}, \{c_3\}, ..., \{c_{n-1}, c_n\}]$, then

$$\pi(\mathbf{A}_c) = [\{\mathbf{a}_c^1, \mathbf{a}_c^2\}, \{\mathbf{a}_c^3\}, ..., \{\mathbf{a}_c^{n-1}, \mathbf{a}_c^n\}]. \quad (3)$$

In this way, an attention vector sequence is divided into several subsequences and each subsequence represents the attention of one word. 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 $\{\mathbf{a}_c^s, ..., \mathbf{a}_c^{s+l-1}\}$ into one attention vector $\mathbf{a}_w^i$ for w_i with the mixed pooling strategy (Yu et al., 2014) [3]. Then we execute the piecewise upsampling operation over each $\mathbf{a}_w^i$ to keep input and output dimensions unchanged for the sake of plug and play. The detailed process can be summarized as:

$$\mathbf{a}_w^i = \lambda \text{ Maxpooling}(\{\mathbf{a}_c^s, ..., \mathbf{a}_c^{s+l-1}\}) + (1-\lambda) \text{ Meanpooling}(\{\mathbf{a}_c^s, ..., \mathbf{a}_c^{s+l-1}\}), \quad (4)$$

$$\hat{\mathbf{A}}_c[s : s+l-1] = \mathbf{e}_l \otimes \mathbf{a}_w^i, \quad (5)$$

where $\lambda \in R^1$ is a weighting trainable variable to balance the mean and max pooling, $\mathbf{e}_l = [1, ..., 1]^T$ represents a l-dimensional all-ones vector, l is the length of w_i, $\mathbf{e}_l \otimes \mathbf{a}_w^i = [\mathbf{a}_w^i, ..., \mathbf{a}_w^i]$ denotes the kronecker product operation between $\mathbf{e}_l$ and $\mathbf{a}_w^i$, $\hat{\mathbf{A}}_c \in \mathbb{R}^{n \times n}$ 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:

$$\hat{\mathbf{H}} = \hat{\mathbf{A}}_c \mathbf{V} \mathbf{W}_v, \quad (6)$$

where $\mathbf{V} = \mathbf{H}$, $\mathbf{W}_v \in \mathbb{R}^{d \times d}$ 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 $\hat{\mathbf{A}}_c^k (1 \le k \le K)$ and corresponding representation $\hat{\mathbf{H}}^k$. With multi-head attention architecture, the output can be expressed as follows:

$$\overline{\mathbf{H}} = \text{Concat}(\hat{\mathbf{H}}^1, \hat{\mathbf{H}}^2, ..., \hat{\mathbf{H}}^K)\mathbf{W}_o. \quad (7)$$

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 π. 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 π 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 $\overline{\mathbf{H}}^1, ..., \overline{\mathbf{H}}^M$ by Eq. 7. Then we propose to fuse these semantically different representations as follows:

$$\tilde{\mathbf{H}} = \sum_{m=1}^{M} \tanh(\overline{\mathbf{H}}^m \mathbf{W}_g), \quad (8)$$

where $\mathbf{W}_g$ is a parameter matrix and $\tilde{\mathbf{H}}$ 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. For detailed hyper-parameter settings, please see Appendix A.3. Besides, three popular CWS tools: thulac (Sun et al., 2016), ictclas (Zhang et al., 2003) and hanlp (He, 2014) are employed to segment sequence.

We carried out experiments on five Chinese NLP tasks, including sentiment classification (EC), named entity recognition (NER), sentence pair matching (SPM), natural language inference (NLI) and machine reading comprehension (MRC). Statistics of the above tasks and the corresponding datasets are detailed in Appendix A.2.

3.2 Experiment Results

Table 2 shows the performances on five classical Chinese NLP tasks with six public datasets. Gener-

[3] Other pooling methods such as max pooling or mean pooling also works. Here we choose mixed pooling because it has the advantages of distilling the global and the most prominent features in one word at the same time.

Task	EC		NER	SPM	NLI	MRC	
Dataset	ChnSenti	weibo-100k	Ontonotes	LCQMC	XNLI	DRCD [EM]	[F1]
BERT	94.72	97.31	79.18	86.50	78.19	85.57	91.16
+MWA	95.34(+0.62)	98.14(+0.83)	79.86(+0.68)	86.92(+0.42)	78.42(+0.23)	86.86(+1.29)	92.22(+1.06)
BERT-wwm	94.38	97.36	79.28	86.11	77.92	84.11	90.46
+MWA	95.01(+0.63)	98.13(+0.77)	**80.32**(+1.04)	86.28(+0.17)	78.68(+0.76)	87.00(+2.89)	92.21(+1.75)
ERNIE	95.17	97.30	77.74	87.27	78.04	87.85	92.85
+MWA	**95.52**(+0.35)	**98.18**(+0.88)	78.78(+1.04)	**88.73**(+1.46)	**78.71**(+0.67)	**88.61**(+0.76)	**93.72**(+0.87)

Table 2: Evaluation results regarding each model on different datasets. Bold marks highest number among all models. 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 ($p < 0.05$). Numbers in brackets indicate the absolute increase over baseline models.

ally, our method consistently outperforms all baselines on all of five tasks, which demonstrates the effectiveness and universality of the proposed approach. Moreover, the Wilcoxon's test shows that significant difference ($p < 0.05$) exits between our model and baseline models.

In detail, on the two datasets of EC 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 a better guidance to label each character, which is consistent with the conclusion in (Zhang and Yang, 2018). Similarly, 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 the 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 a strong baseline by introduce a character-level multi-head self-attention layer (CA) to exclude the benefits from additional parameters. As shown in Table 3, adding additional parameters by introducing extra CA layer can benefit the PLMs. Compared with character-level attention, 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 employing multiple segmenters and fusing them together could introduce richer segmentation information and further improve the performance.

Model	BERT	BERT-wwm	ERNIE
Original	92.06	91.68	92.61
+CA	92.37	92.22	93.42
+WA$_{thulac}$	92.84	92.73	93.89
+WA$_{ictclas}$	93.05	92.90	93.75
+WA$_{hanlp}$	92.91	93.21	93.91
+MWA	**93.59**	**93.72**	**94.21**

Table 3: F1 results of ablation experiments on the DRCD dev set. Scores are the average of five models.

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 be also applied to English PLMs to bridge the semantic gap between the whole word and the segmented WordPiece tokens, which we leave for future work.

References

Xuejun Bai, Guoli Yan, Simon P Liversedge, Chuanli Zang, and Keith Rayner. 2008. Reading spaced and unspaced chinese text: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5):1277.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019. Cnn-based chinese ner with lexicon rethinking. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4982–4988. AAAI Press.

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.

Yanzeng Li, Tingwen Liu, Diying Li, Quangang Li, Jinqiao Shi, and Yanqiu Wang. 2018. Character-based bilstm-crf incorporating pos and dictionaries for chinese opinion target extraction. In *Asian Conference on Machine Learning*, pages 518–533.

Hongxia Meng, Xuejun Bai, Chuanli Zang, and Guoli Yan. 2014. Landing position effects of coordinate and attributive structure compound words. *Acta Psychologica Sinica*, 46(1):36–49.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. Drcd: a chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920*.

Maosong Sun, Xinxiong Chen, Kaixu Zhang, Zhipeng Guo, and Zhiyuan Liu. 2016. Thulac: An efficient lexical analyzer for chinese. Technical report.

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*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2017. Exploiting word internal structures for generic Chinese sentence representation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 298–303, Copenhagen, Denmark. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Dingjun Yu, Hanli Wang, Peiqiu Chen, and Zhihua Wei. 2014. Mixed pooling for convolutional neural networks. In *International Conference on Rough Sets and Knowledge Technology*, pages 364–375. Springer.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. Hhmm-based chinese lexical analyzer ictclas. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 184–187. Association for Computational Linguistics.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564.