A Joint Multiple Criteria Model in Transfer Learning for Cross-domain Chinese Word Segmentation

Kaiyu Huang, Degen Huang*, Zhuang Liu and Fengran Mo
School of Computer Science, Dalian University of Technology
{kaiyuhuang,zhuangliu,fengranmo}@mail.dlut.edu.cn
huangdg@dlut.edu.cn

Abstract

Word-level information is important in natural language processing (NLP), especially for the Chinese language due to its high linguistic complexity. Chinese word segmentation (CWS) is an essential task for Chinese downstream NLP tasks. Existing methods have already achieved a competitive performance for CWS on large-scale annotated corpora. However, the accuracy of the method will drop dramatically when it handles an unsegmented text with lots of out-of-vocabulary (OOV) words. In addition, there are many different segmentation criteria for addressing different requirements of downstream NLP tasks. Excessive amounts of models with saving different criteria will generate the explosive growth of the total parameters. To this end, we propose a joint multiple criteria model that shares all parameters to integrate different segmentation criteria into one model. Besides, we utilize a transfer learning method to improve the performance of OOV words. Our proposed method is evaluated by designing comprehensive experiments on multiple benchmark datasets (e.g., Bakeoff 2005, Bakeoff 2008 and SIGHAN 2010). Our method achieves the state-of-the-art performances on all datasets. Importantly, our method also shows a competitive practicability and generalization ability for the CWS task.

Corpora	Zhang	Xiao	Fan	attend	a tournament	
PKU	张	小凡		参加	比武	大会
MSRA	张小凡			参加	比武大会	
Zhuxian	张小凡			参加	比武	大会

Table 1: Illustration of different segmentation criteria on three popular datasets

1 Introduction

In the extensive researches on natural language processing (NLP), most of the tasks are based on word-level methods because word is the smallest linguistic unit in natural languages. It has rich feature information. However, the situation is totally different when dealing with the Chinese language. There is not clearly delimiter between Chinese words, instead the blank space is regarded as a delimiter in most western languages. Different segmentation results may lead to different feature information. Thus, Chinese word segmentation (CWS) is an essential task, which will significantly affect the effectiveness of downstream Chinese NLP tasks. Recently, the approaches for CWS have already achieved a good performance in large-scale annotated corpora, as reported by related researches (Huang and Zhao, 2007; Zhao et al., 2019). Most of the effective approaches fall into two major research fields: the statistical machine learning method and the neural network method. The former is mainly based on Conditional Random Fields (CRF), which is considered as the most effective statistical machine learning method for CWS (Zhao and Kit, 2008; Zhao et al., 2010). However, the statistical machine learning method always heavily relies on hand-craft features. To minimize the efforts in feature engineering, more and more researches are focus on neural network method (Pei et al., 2014; Chen et al., 2015a,b). Furthermore, following the rapid development of neural network models, variations on neural network methods for CWS have already gained comparable results as the state-of-the-art statistical machine learning techniques (Cai et al., 2017; Zhou et al., 2017; Ma et al., 2018; Meng et al., 2019).

Nevertheless, there are still two important issues on the CWS task. One important issue is that almost all effective methods are limited by large-scale annotated corpora, these methods will lead to a weak generalization ability. The results may decline rapidly when the methods deal with

[*] Corresponding author

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 3873–3882,
November 16–20, 2020. ©2020 Association for Computational Linguistics

a cross-domain situation. Since there are many out-of-vocabulary (OOV) words in cross-domain scenarios, and the character feature information is different in another unrelated domain. For example, the Chinese character “莫(Mo/not to do)” is always trained as a surname of Chinese people in most domains, especially when a slice “莫言(Mo Yan/not to say)” appears in a sentence, it probably should be a Chinese person who won the Nobel Prize for literature. However, the situation is totally different in the Chinese novel. In Chinese novel domain, “莫(Mo/not to do)” always means “not to do something”. When the slice “莫言(Mo Yan/not to say)” appears in a famous Chinese novel, the meaning of “莫言(Mo Yan/not to say)” is “do/does not to say” definitely. The current methods hardly segment it correctly because of the low generalization ability and robustness. The other important issue is that Chinese word segmentation criterion is multiple, and most novel methods depend on large-scale corpora. If the large-scale corpora have different criteria, which are shown in Table 1, the method that is trained by a heterogeneous criterion corpus is hard to segment correctly. In the previous researches, the usual solutions are to train different models to adapt to multiple segmentation criteria.

In this paper, we propose a joint multiple criteria method for both standard and cross-domain simplified Chinese word segmentation. The method utilizes a novel pre-trained (RoBERTa-WWM) model (Cui et al., 2019), which adequately trained a rich Chinese character vector embedding. With the richness of the pre-trained model, our method for CWS can obtain a robust generalization ability to deal with the cross-domain situation. In order to further improve the performance of the model, we consider improving the amount of training data through the process of transfer learning. We adopt a strategy that integrates several different segmentation criteria into a single straightforward model. The benefit is that we do not need many models to fit multiple segmentation criteria, and the method improves the amount of training data in disguised simultaneously.

To sum up, the contributions of this paper are as follows.

- We present a straightforward transfer learning method based on RoBERTa to solve CWS problems mentioned above, and make use of the rich pre-trained model that extracted abundant feature information and linguistic context, making the word-formation ability of the model strong. The method achieves state-of-the art performance on in-domain and cross-domain CWS benchmarks.
- There is a large number of parameters in the RoBERTa-based model. We share all the parameters without complex neural network architectures in the training step. It can control the explosive growth of the total parameters while improves the performances on several datasets for CWS.
- Our proposed method is straightforward and effective. We do not need to devise much manual information such as lexicon, n-gram feature, and statistical information. It matches with the benefits of neural network properly.

2 Related Work

Since Xue (2003) first formalizes CWS task as a sequence labeling problem, many researches depending on supervised machine learning methods have already achieved good performance for CWS. Peng et al. (2004) utilized the CRF methods for CWS, since then CRF became the most popular machine learning method for CWS task. Variations of CRF based model achieved good performances for CWS (Tseng et al., 2005; Zhao and Kit, 2008; Zhao et al., 2010; Sun et al., 2012; Zhang et al., 2013). With the development of neural network, more and more researchers made gradual progress with a wide range of neural methods (Zheng et al., 2013; Pei et al., 2014; Chen et al., 2015a,b; Cai and Zhao, 2016; Cai et al., 2017), and the performances on neural methods have already approximated state-of-the-art performances on statistical methods.

Neural network methods can incorporate the information in the model easily and effectively by means of automatic feature extraction. Thus, it is possible to train multiple segmentation criteria into a single model well with neural network methods while it is a challenge on previous statistical methods. Chen et al. (2017) was first to propose a multi-criteria learning method for CWS, in which the method adopted shared layers and private layers. However, there is still a gap with independent segmentation criterion method. He et al. (2018) improved the performance on the same base Bi-LSTM (bidirectional Long Short-Term Memory Network) (Graves and Schmidhuber, 2005) model, it adopted a simple and effective method to integrate differ-

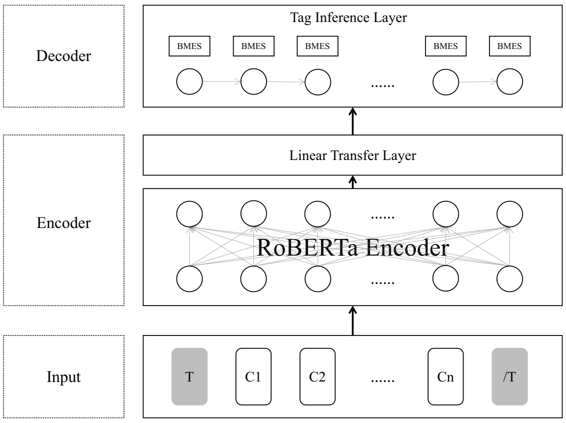

Figure 1: An overview of proposed model architecture.

ent segmentation criteria by adding two tags in a sentence (Johnson et al., 2017). Qiu et al. (2019) changed the base model to Transformer (Vaswani et al., 2017), and Huang et al. (2019) utilized BERT (Devlin et al., 2018) that is based on Transformer to extract feature information. Both of these two methods achieved the state-of-the-art performances on benchmark datasets.

The approaches mentioned above mainly focus on in-domain benchmarks, and there is still much room for improvement using the neural network method (Huang et al., 2017; Zhang et al., 2018; Zhao et al., 2018; Ye et al., 2019). Most of these methods leverage external resources to alleviate the OOV issue. Our proposed method is inspired by this thought, and uses rich pre-trained embeddings to relieve the weakness in cross-domain. The method not only solves the multiple segmentation criteria issue with a straightforward architecture, but also solves the cross-domain CWS problem. With the help of the pre-trained embedding, the transfer learning method does not need to learn from scratch, and has a robust generalization ability.

3 Model Architecture

Figure 1 shows our proposed model architecture which is quite brief. We do not pay attention to complicating the neural network. Meanwhile, the strategy of integrating criteria is first proposed on Johnson et al. (2017) for machine translation transfers multiple segmentation criteria into one model with minimal effort.

3.1 Encode Layer

According to Ma et al. (2018), the complexity of a neural network for CWS can hardly affect the performance since the CWS is a task on the superficial linguistic representation. The features of the characters are shallow. There will be a competitive performance on simple neural network architecture. The real factor that leads to the gap of CWS task is under-training instead of bad-training. Thus, we utilize the Whole Word Masking RoBERTa model, which is trained by large unlabeled Chinese data.

The input of encode layer consists of three parts that are token embedding $\mathbf{E}_t$, position embedding $\mathbf{E}_p$ and segment embedding $\mathbf{E}_s$. Given a character sentence $C = \{C_1C_2C_3...C_{n-2}C_{n-1}C_n\}$ as the input. The position sequence of the input is $P = \{P_1P_2P_3...P_{n-2}P_{n-1}P_n\}$. The sequence is converted to a vector matrix $\mathbf{E}_t$. P is also mapped into a feature matrix $\mathbf{E}_p$. Because of the specificity for CWS, all segment embeddings of the sequences are regarded as the same mapping relation $\mathbf{E}_s$. The input is

$$\mathbf{E}_{input} = E_t + E_p + E_s \tag{1}$$

The encode layer consists of several transformer encoders(Vaswani et al., 2017), and it is bidirectional. The transformer encoder utilizes several multi-head self-attention layers to extract the contextual feature for each character. The multi-head self-attention layer adopts “Scaled Dot-Product Attention” to compute representation. The “Scaled Dot-Product Attention” function is:

$$Attention\left(Q, K, V\right) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where Q, K, V represents a query and a set of key-value pairs through a linear transformation respectively, and d_k is the dimension of K.

Instead of performing a single attention function, the multi-head self-attention layer can extract contextual features from different representation spaces. Given an input sequence of vector $E_{input} \in \mathbb{R}^{L*d_{model}}$, where L is the length of the sequence, and d_{model} is the dimension of it. A multi-head self-attention layer is:

$$MultiHead\left(E_{input}\right) = \left[head_1, ..., head_k\right]W^O \tag{3}$$

$$head_i = Attention\left(E_{input}W_i^Q, E_{input}W_i^K, E_{input}W_i^V\right) \tag{4}$$

where W^O, W_i^Q, W_i^K, W_i^V are trainable parameters. And a layer normalization is adopted in the end of each multi-head self-attention layer.

Corpora		PKU	MSRA	CTB	CNC	UDC	SXU	ZX	Cross-domain			
									L	C	M	F
Words	Train	1.1M	2.4M	0.7M	6.6M	0.1M	0.5M	88K	-	-	-	-
	Test	0.1M	0.1M	52K	0.7M	12K	0.1M	34K	35K	35K	31K	33K
Chars	Train	1.8M	4.0M	1.2M	10M	0.2M	0.8M	0.1M	-	-	-	-
	Test	0.2M	0.2M	86K	1.1M	18K	0.2M	48K	50K	54K	51K	53K
OOV rate (%)		3.5	2.1	5.0	0.7	10.6	4.6	5.5	6.2	6.8	9.0	3.8

Table 2: The sizes of different benchmark datasets. "L" represents literature domain. "C" represents computer domain. "M" represents medicine domain. "F" represents finance domain.

With the rich pre-trained model, the trainable parameters have already been covered with a wealth of character information in the hidden states. These feature information could compensate for the weaknesses of unknown characters. To adapt the pre-trained model to CWS, we utilize a linear transfer layer to integrate the hidden states into CWS.

3.2 Multiple Criteria on Solution

There are many parameters in the pre-trained model, so it would be impractical to train different models for dealing with different criteria and domains. Inspired by the similarity with the method, it needs to integrate different languages into one model. Our proposed strategy considers criteria as languages. The straightforward and effective method is that each of the input sentences attaches a pair of tag identifiers $< tag > and < /tag >$ at the beginning and end of the sentence respectively. tag represents the specific criterion or domain. For instance, if an input sentence C belongs to "PKU" datasets, the sentence changes into $< pku > C < /pku >$ as the input of encode layer. These specific identifiers can affect the contextual representation within the scope of the sentence. It is similar to domain-aware protection, making the correct decision matching criterion or domain of the unsegmented sentence. We do not pay attention to producing external computing. And it will save much room for creating model architectures.

3.3 Tag Inference

According to Xue (2003), our proposed method also converts CWS task to a character based sequence labeling problem. One commonly used labeling set is a 4-tag set $T = \{B, M, E, S\}$, representing the **b**egin, **m**iddle, **e**nd of a word, or a **s**ingle character forming a word. The aim of the character based sequence labeling task is to find the most possible path of $Y^* = \{Y_1 Y_2 ... Y_{n-1} Y_n\}$:

$$Y^* = \underset{Y \in T^n}{argmax}\, p\,(Y|X) \qquad (5)$$

hidden state size	768
optimizer	Bert Adam
learning rate	uniform-float[1e-5,**2e-5**,1e-4]
batch size	uniform-integer[16,32,**64**]
dropout	uniform-float[**0.1**,0.2,0.3,0.5]
epochs	15

Table 3: The hyper-parameters settings, the best assignments are highlighted.

In the previous methods, many researchers adopted a CRF decode layer to improve the performance for sequence labeling task (Lample et al., 2016). In particular, the core algorithm of the neural CRF layer is a transition matrix during the decode step. The transition matrix can learn constraint rules between two tags in order to enhance accuracy. It is effective for most complicated NLP tasks. However, the ability of improving accuracy is limited by utilizing the CRF layer because there is a high accuracy of each character tag itself on our model. Meanwhile CRF layer will have larger time complexity and space complexity. So we utilize a lightweight decode layer $Softmax$, which increases smaller time complexity. The loss function is cross-entropy:

$$Loss\,(y, y^*) = -\sum_{x} y(x) log y^*(x) \qquad (6)$$

where y denotes the gold sequence labeling, y^* denotes the output of decode layer.

4 Experiment

4.1 Datasets and Experimental Setup

For verifying the high performance of the joint multiple criteria model in transfer learning for in-domain and cross-domain CWS, we do comparative experiments on several simplified Chinese

Models		PKU	MSRA	CTB	CNC	UDC	SXU	ZX	Avg.In4
Single criterion learning									
Chen et al. (2017)	F	93.30	95.84	95.30	-	-	95.17	-	94.09
	R_{oov}	66.09	66.28	76.47	-	-	71.27	-	70.02
He et al. (2018)	F	95.22	97.29	96.27	97.11	93.98	95.80	95.57	96.15
	R_{oov}	-	-	-	-	-	-	-	-
Gong et al. (2019)	F	95.74	96.46	97.09	-	-	95.18	-	96.12
	R_{oov}	72.70	69.90	81.80	-	-	69.69	-	73.52
Qiu et al. (2019)	F	96.39	98.07	96.43	-	-	97.08	-	96.70
	R_{oov}	72.82	73.75	82.82	-	-	77.95	-	76.84
Ours	F	96.67	98.12	**97.56**	**97.26**	**97.86**	97.52	**96.77**	97.47
	R_{oov}	79.13	80.65	**88.24**	58.05	**92.58**	85.01	**86.12**	83.26
Multiple criteria learning									
Chen et al. (2017)	F	94.32	96.04	96.18	-	-	96.04	-	95.65
	R_{oov}	72.67	71.60	82.48	-	-	77.10	-	75.96
He et al. (2018)	F	96.06	97.25	96.70	97.00	94.44	96.47	95.72	96.62
	R_{oov}	-	-	-	-	-	-	-	-
Gong et al. (2019)	F	96.15	97.78	97.26	-	-	97.25	-	97.11
	R_{oov}	69.88	64.20	83.89	-	-	78.69	-	74.17
Qiu et al. (2019)	F	96.41	98.05	96.99	-	-	97.61	-	97.27
	R_{oov}	78.91	78.92	87.00	-	-	85.08	-	82.48
Ours	F	**96.85**	**98.29**	**97.56**	97.19	97.69	**97.56**	96.46	**97.56**
	R_{oov}	**82.35**	**81.75**	88.02	**59.44**	91.40	**85.73**	82.51	**84.46**

Table 4: The results on test data of 7 standard CWS datasets. Here, F and R_{oov} represent F1 value and the recall of OOV words respectively. "Avg.In4" is the average of PKU, MSRA, CTB and SXU. The maximum values of evaluation are highlighted for each dataset.

datasets, including Bakeoff 2005, Bakeoff 2008, SIGHAN 2010 (cross-domain datasets), and other open datasets. The sizes of corpora are shown in Table 2. We randomly pick 10% sentences from the training data as the development data for model tuning. Similar to a previous paper (Cai et al., 2017), we convert all digits, punctuation, and Latin letters to half-width, dealing with the full/half-width mismatch between training and test data. The continuous Latin characters and digits are generalized to a unique token. Note that there is no training data for the cross-domain datasets, so the tag of cross-domain datasets is set to PKU which is the most similar to them.

4.2 Multiple Criteria Result

We follow the majority of hyper-parameters of the original RoBERTa-WWM model, adjusting a few crucial hyper-parameters. The hypter-parameters and search ranges that are shown in Table 3. We deploy the model on GPU(Nvidia Tesla K40c). One epoch with 1.7M tokens costs about 6 hours in the training step. Our implementation is based on Pytorch (Wolf et al., 2019; Paszke et al., 2019), a dynamic neural graph framework for deep learning.[1]

Table 4 shows the results of both single criterion method and multiple criteria method on several benchmark datasets.

We first compare our method with the previous popular methods. Three of them are based on LSTM neural architecture (Chen et al., 2017; He et al., 2018; Gong et al., 2019). Our method and Qiu et al. (2019) are based on Transformer. It is observed that the performance on Transformer is better than it on LSTM from the table. In particular, our method utilizes the pre-trained embedding that is more effective on all seven simplified Chinese benchmark datasets. A full-training language model can improve the generalization ability and robustness of model.

Furthermore, our proposed method adopts a strategy to integrate all training datasets into one model.

[1] Our code are available at `https://github.com/koukaiu/dlut-nihao`

Models		Literature	Computer	Medicine	Finance	Avg.
non-DL	Huang and Tong (2012)	94.66	95.36	94.69	96.26	95.24
	Liu et al. (2014)	92.49	94.07	92.63	95.54	93.68
	$SOTA_p$	95.5	95.0	93.8	96.0	95.08
DL	Chen et al. (2015b)	92.89	93.71	92.16	95.20	93.49
	Cai et al. (2017)	92.90	94.04	92.10	95.38	93.61
	Huang et al. (2017)	94.33	93.99	92.26	95.81	94.10
	Zhao et al. (2018)	93.23	95.32	93.73	95.84	94.53
	Zhang et al. (2018)	94.76	94.70	94.18	96.06	94.93
	Baseline	93.13	93.19	91.73	94.96	93.25
	+pre-trained	94.96	94.86	94.23	96.33	95.10
	Ours	**96.13**	**96.08**	**95.21**	**96.82**	**96.06**

Table 5: The F1 values on test data of SIGHAN 2010 cross-domain datasets. Here, "$SOTA_p$" represents the previous maximum F1 values on SIGHAN 2010 open test task of each domain, including three results (Computer, Medicine, Finance) from Gao and Vogel (2010) and one result (Literature) from Huang et al. (2010). The currently maximum values of evaluation are highlighted for each domain dataset.

Models	P	R	F	R_{oov}
Ours	**97.27**	**96.43**	**96.85**	82.35
Baseline	95.44	94.96	95.20	62.82
+second hidden	96.40	95.34	95.87	77.89
+second-to-last hidden	96.31	95.31	95.81	80.04
+sum last four hidden	96.09	95.36	95.73	81.37
+sum all 12 hidden	96.31	95.10	95.70	**82.66**

Table 6: The results by adopting differnt layers of the pre-trained model on PKU. Here "P" is the precision, "R" is the recall, "F" is the F1-value and "R_{oov}" is the recall of OOV words.

It not only reduces the sum of parameters by N (the number of different segmentation criteria) times, but also improves the performances slightly on four (PKU, MSRA, CTB, SXU) of seven datasets. The most important thing is that the knowledge of multiple segmentation criteria is merged together by our method. We also compared some open datasets with He et al. (2018). Our proposed method has a significant improvement compared to the previous works. Note that the R_{oov} of CNC is relatively lower than others. One possible reason is that the training set of CNC is extensive, the OOV words are almost the unconventional words. It is challenging to segment them on current technology. The other possible reason is that there are some errors in the corpus itself.

4.3 Cross-domain Result

We compare our model with the previous effective methods for cross-domain CWS, shown in Table 5. No matching development set is provided for the cross-domain datasets, so we follow hyper-parameters of PKU set. Both of statistical method (non-DL) and neural network method (DL) have competitive performances on cross-domain datasets. However, according to the results in Table 5, it is observed that neural CWS methods fall short of the performances compared with statistical methods in the previous works. With external resources, some neural CWS methods are close to the previous state-of-the-art performances for cross-domain CWS (Zhao et al., 2018; Zhang et al., 2018). For verifying the contribution of the pre-trained model, we adopt a popular neural architecture (Bi-LSTM) as the baseline model, and utilize the pre-trained embedding based on the baseline model to improve the performance. The difference between these two methods reflects the role of pre-trained embedding partly. It effectively alleviates the OOV issue by using rich pre-trained embedding instead of modifying the model architecture. From the results, the pre-trained method has already achieved the best performance of statistical methods. It supplies a gap on a pure neural CWS model that does not utilize any external resources. Our proposed transfer learning method not only takes full advantages of pre-trained embedding, but also adopts a strategy to increase the scale of training sample in disguise. As we know, the scale of training samples is the key to improve the performance with neural methods. Our method has achieved state-of-the-art performance compared with the previous non-DL and DL methods on all of four cross-domain SIGHAN 2010 datasets.

Methods	PKU	MSRA
Zhao et al. (2010)	96.7	98.0
Cai et al. (2017)	95.8	97.1
Yang et al. (2017)	96.3	97.5
Zhou et al. (2017)	97.8	96.0
Ma et al. (2018)	96.1	97.4
Huang et al. (2019)	96.6	97.9
Meng et al. (2019)	96.7	**98.3**
Ours	**96.9**	**98.3**

Table 7: The F1-values on PKU and MSRA Bakeoff 2005 datasets. The maximum values of evaluation are highlighted for each column.

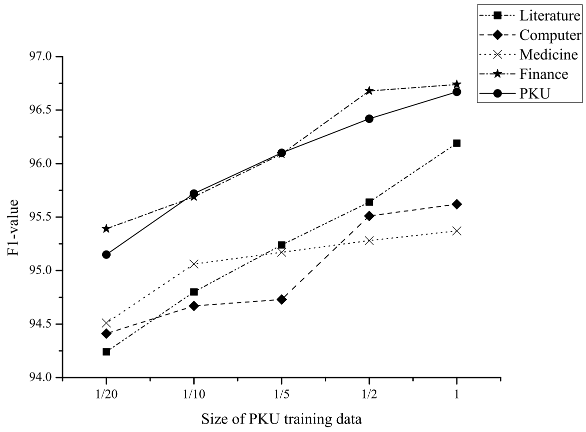

Figure 2: The F1-values of our method on PKU and cross-domain SIGHAN 2010 datasets. The X-axis represents the size of training set, the Y-axis represents the the F1-values. The icons of different datasets are described at the top right of the figure.

4.4 Generalization Ability

Most of the present neural CWS methods adopt the pre-trained embedding to avoid the OOV problem. To varying degrees, the pre-trained embedding improves the performances of the neural CWS method. In other words, how to utilize the pre-trained embedding is a key to enhance the generalization ability of a neural CWS method. We adopt several types of pre-trained embeddings, the results are shown in Table 6. Indeed, the different pre-trained embeddings improve the performance of the baseline model. However, our method that utilizes the concept of transfer learning improves the generalization ability more forcefully. In particular, the performances of our method on PKU and MSRA benchmark datasets are state-of-the-art, shown in Table 7.

Another influencing factor of a supervised neural method is the size of the training set. In particular, the Transformer needs a large size of the training set more than other previous neural architectures. We utilize different sizes of PKU training data to evaluate the performances on 5 datasets that include four cross-domain datasets and a PKU benchmark dataset, shown in Figure 2. It is observed that there is a better performance on a larger size of the training set. In other words, we can enhance the generalization ability by adding training data. Under the premise of not adding manual annotation, we might utilize the multiple criteria available to make the model more robust. According to the results of several experiments, our method shows a competitive practicability and generalization ability.

4.5 Error Analysis

In order to guide future research directions for Chinese word segmentation, we analyze three typical types of errors in our method by manual and non-manual.

The first one is that there are many errors due to annotation inconsistency or annotation errors. For instance, the word "操作系统(operating system)" occurs nine times annotated as "操作(operate)+系统(system)" and more than ten times as "操作系统(operating system)" in the same context. There are many similar situations in the corpora through the consistency checking. Besides, "国故(national cultural heritage)" should be regarded as a word that is even difficult for a Chinese to understand. In the context "他通过整理国故而帮助建立了学说"(He established a doctrine by concluding national cultural heritage), the gold result is given as "他(He)/通过(pass by)/整理(conclude)/国(nation)/故(heritage)/而(while)/帮助(help)/建立(establish)/了(an empty word)/学说(doctrine)". Furthermore, the word "国故(national cultural heritage)" is separated into two single words. Words that are difficult to understand are high probability wrong in gold results. Unfortunately, These errors come from the original corpus itself, so we argue that it is not an algorithm problem. It might be proceeded with amending the corpus.

The second one is that the model hesitates when a prefix/suffix might be an independent single word. For instance, "案(file)" is a suffix word, frequently appeared in a word together with another two characters like "修正案(amendment)" and "走私案(smuggling case)". When the model predicts "犯罪(crime)案(case)", it is great probability to merge them together incorrectly. Similarly, the prefix/suffix problem is trapped in the issue of consistency commonly.

The last one is that the performances of the longer OOV words are unsatisfactory. In particular, some longer personal names that do not contain clearly feature information are hard to segment. Unlike "约翰大卫(John David)" and "柴可夫斯基(Tchaikovsky)" that are obviously treated as an English name and a Russian name, the sequence of "山(mountain)鹿(deer)素(element)行(walk)" is segmented as four single words, while it is a Japanese researcher (Yamaga Sokou). In addition, the errors not only limit to personal names, but also distinguish the word boundary incorrectly. Should "国营企业(state-owned enterprise)" be segmented as one word or two words "国营(state-owned)+企业(enterprise)"? It is hard to segment correctly for human, the model absolutely struggles to distinguish the boundary.

5 Conclusion

In this paper, we construct a transfer layer structure that leverages the pre-trained feature information for CWS and exploit a transfer tag to boost joint multiple criteria learning. The model could relieve the OOV problem for Chinese word segmentation and achieves the best performance in comparison with state-of-the-art techniques for both in-domain and out-of-domain Chinese word segmentation. Extensive experiments on seven in-domain and four cross-domain datasets for Chinese word segmentation confirm the superiority of our model over all other advanced methods. In summary, the advantages of our model are twofold. First, the model has a stronger robustness with a straightforward transfer learning method. The performance of our model is better, especially when dealing with high OOV rate data. Second, our model effectively solves the parameters exploding due to different segmentation criteria. We do not need to design any redundant structures. Nevertheless, there is still a gap in a real-word situation. In the future, we will continue studying the efficiency of the neural architecture, and pay attention to improving the speed of both training and testing steps on an ever-increasing dataset. In particular, we will enhance the practicability of Chinese word segmentation to improve the effectiveness of other downstream Chinese NLP tasks.

Acknowledgments

We would like to thank the reviewers for their helpful comments and suggestions to improve the quality of the paper. The authors gratefully acknowledge the financial support provided by the National Natural Science Foundation of China under (No.U1936109, 61672127).

References

Deng Cai and Hai Zhao. 2016. Neural word segmentation learning for chinese. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 409–420, Berlin, Germany. Association for Computational Linguistics.

Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu, and Feiyue Huang. 2017. Fast and accurate neural word segmentation for chinese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 608–615, Vancouver, Canada. Association for Computational Linguistics.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, and Xuanjing Huang. 2015a. Gated recursive neural network for chinese word segmentation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1744–1753, Beijing, China. Association for Computational Linguistics.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Pengfei Liu, and Xuanjing Huang. 2015b. Long short-term memory neural networks for chinese word segmentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1197–1206, Lisbon, Portugal. Association for Computational Linguistics.

Xinchi Chen, Zhan Shi, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-criteria learning for chinese word segmentation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1193–1203, Vancouver, Canada. 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. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qin Gao and Stephan Vogel. 2010. A multi-layer chinese word segmentation system optimized for out-of-domain tasks. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*.

Jingjing Gong, Xinchi Chen, Tao Gui, and Xipeng Qiu. 2019. Switch-lstms for multi-criteria chinese word segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6457–6464.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Han He, Lei Wu, Hua Yan, Zhimin Gao, Yi Feng, and George Townsend. 2018. Effective neural solution for multi-criteria word segmentation. In *Smart Intelligent Computing and Applications: Proceedings of the Second International Conference on SCI 2018*, volume 2, page 133. Springer.

Chang-Ning Huang and Hai Zhao. 2007. Chinese word segmentation: a decade review. *Journal of Chinese Information Processing*, 21(3):8–19.

Degen Huang and Deqin Tong. 2012. Context information and fragments based cross-domain word segmentation. *China Communications*, 9(3):49–57.

Degen Huang, Deqin Tong, and Yanyan Luo. 2010. Hmm revises low marginal probability by crf for chinese word segmentation. In *CIPS-SIGHAN Joint Conference on Chinese Language Processing*.

Shen Huang, Xu Sun, and Houfeng Wang. 2017. Addressing domain adaptation for chinese word segmentation with global recurrent structure. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 184–193.

Weipeng Huang, Xingyi Cheng, Kunlong Chen, Taifeng Wang, and Wei Chu. 2019. Toward fast and accurate neural chinese word segmentation with multi-criteria learning. *arXiv preprint arXiv:1903.04190*.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

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 the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Yijia Liu, Yue Zhang, Wanxiang Che, Ting Liu, and Fan Wu. 2014. Domain adaptation for crf-based chinese word segmentation using free annotations. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 864–874, Doha, Qatar. Association for Computational Linguistics.

Ji Ma, Kuzman Ganchev, and David Weiss. 2018. State-of-the-art chinese word segmentation with bi-lstms. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4902–4908.

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. In *Advances in Neural Information Processing Systems*, pages 2742–2753.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Wenzhe Pei, Tao Ge, and Baobao Chang. 2014. Max-margin tensor neural network for chinese word segmentation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 293–303, Baltimore, Maryland, USA. Association for Computational Linguistics.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING '04)*, pages 562–568, Geneva, Switzerland. Association for Computational Linguistics, Association for Computational Linguistics.

Xipeng Qiu, Hengzhi Pei, Hang Yan, and Xuanjing Huang. 2019. Multi-criteria chinese word segmentation with transformer. *arXiv preprint arXiv:1906.12035*.

Xu Sun, Houfeng Wang, and Wenjie Li. 2012. Fast online training with frequency-adaptive learning rates for chinese word segmentation and new word detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 253–262, Jeju, Republic of Korea. Association for Computational Linguistics.

Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter for sighan bakeoff2005. In *Proceedings of the Fourth SIGHAN workshop on Chinese Language Processing*, pages 168–171, Jeju Island, Korea.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Nianwen Xue. 2003. Chinese word segmentation as character tagging. *Computational Linguistics and Chinese Language Processing*, 8(1):29–48.

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 (Volume 1: Long Papers)*, pages 839–849, "Vancouver, Canada". "Association for Computational Linguistics".

Yuxiao Ye, Weigang Li, Yue Zhang, Likun Qiu, and Jian Sun. 2019. Improving cross-domain chinese word segmentation with word embeddings. 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 2726–2735.

Longkai Zhang, Houfeng Wang, Xu Sun, and Mairgup Mansur. 2013. Exploring representations from unlabeled data with co-training for chinese word segmentation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 311–321, Seattle, Washington, USA. Association for Computational Linguistics.

Qi Zhang, Xiaoyu Liu, and Jinlan Fu. 2018. Neural networks incorporating dictionaries for chinese word segmentation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Hai Zhao, Deng Cai, Changning Huang, and Chunyu Kit. 2019. Chinese word segmentation: another decade review (2007-2017). *arXiv preprint arXiv:1901.06079*.

Hai Zhao, Chang-Ning Huang, Mu Li, and Bao-Liang Lu. 2010. A unified character-based tagging framework for chinese word segmentation. *ACM Transactions on Asian Language Information Processing*, 9(2):1–32.

Hai Zhao and Chunyu Kit. 2008. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition. In *The Sixth SIGHAN Workshop on Chinese Language Processing*, pages 106–111, Hyderabad, India.

Lujun Zhao, Qi Zhang, Peng Wang, and Xiaoyu Liu. 2018. Neural networks incorporating unlabeled and partially-labeled data for cross-domain chinese word segmentation. In *IJCAI*, pages 4602–4608.

Xiaoqing Zheng, Hanyang Chen, and Tianyu Xu. 2013. Deep learning for chinese word segmentation and pos tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 647–657, Seattle, Washington, USA. Association for Computational Linguistics.

Hao Zhou, Zhenting Yu, Yue Zhang, Shujian Huang, Xinyu Dai, and Jiajun Chen. 2017. Word-context character embeddings for chinese word segmentation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 760–766, Copenhagen, Denmark. Association for Computational Linguistics.