Addressing Domain Adaptation for Chinese Word Segmentation with Global Recurrent Structure

Shen Huang and **Xu Sun** and **Houfeng Wang**[*]
Key Laboratory of Computational Linguistics, Ministry of Education
School of Electronics Engineering and Computer Science, Peking University
Beijing, P.R.China, 100871
huangshenno1,xusun,wanghf@pku.edu.cn

Abstract

Boundary features are widely used in traditional Chinese Word Segmentation (CWS) methods as they can utilize unlabeled data to help improve the Out-of-Vocabulary (OOV) word recognition performance. Although various neural network methods for CWS have achieved performance competitive with state-of-the-art systems, these methods, constrained by the domain and size of the training corpus, do not work well in domain adaptation. In this paper, we propose a novel BLSTM-based neural network model which incorporates a global recurrent structure designed for modeling boundary features dynamically. Experiments show that the proposed structure can effectively boost the performance of Chinese Word Segmentation, especially OOV-Recall, which brings benefits to domain adaptation. We achieved state-of-the-art results on 6 domains of CNKI articles, and competitive results to the best reported on the 4 domains of SIGHAN Bakeoff 2010 data.

1 Introduction

Since Chinese writing system does not have explicit word delimiters, word segmentation becomes an essential first step for further Chinese language processing. In recent years, Chinese Word Segmentation (CWS) has experienced great advancement. One mainstream method is to regard word segmentation task as a sequence labeling problem (Xue, 2003; Peng et al., 2004) where each character is assigned a tag indicating its position in the word. This method has been proved effective as it turns word segmentation into a structured discriminative learning task which can be handled by supervised learning algorithms such as Maximum Entropy (ME) (Berger et al., 1996) and Conditional Random Fields (CRF) (Lafferty et al., 2001). Furthermore, rich features can be incorporated into these systems to improve their performances and most state-of-the-art systems are still based on feature-based models.

Recently, neural network models are drawing increasing attention in Natural Language Processing (NLP) tasks. They significantly reduced feature engineering effort and achieved competitive or state-of-the-art results in many NLP tasks. Collobert et al. (2011) developed a general neural network architecture for sequence labeling tasks. Following this work, many neural network models (Zheng et al., 2013; Pei et al., 2014; Chen et al., 2015a,b) have been applied to CWS and some approached state-of-the-art performance.

However, these neural network models, as well as other supervised methods, do not work well in domain adaptation. In recent years, manually annotated training corpus mostly come from the news domain. When it shifts to other domains such as literature or medicine, where there are many domain-related words that rarely appear in other domains, Out-of-Vocabulary (OOV) word recognition becomes an important problem. Moreover, different domains means different language usages and contexts. Therefore, the In-Vocabulary (IV) word segmentation performance is also affected. As a result, CWS accuracies can drop gravely on cross-domain corpora. For example, consider a sentence "三聚氰胺(melamine) / 致(lead to) / 婴幼儿(baby) / 泌尿系(urinary tract) / 结石(stones)". Here the word "三聚氰胺(melamine)" is a chemical that often appears in medicine-related domains while seldom appears in other domains. It is a four-Chinese-character word

[*]Corresponding author

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 184–193,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

where each character stands for 'three', 'gather', 'cyanide' and 'amine'. The four characters are totally irrelevant. A supervised CWS system trained on news domain corpus would face great challenges on segmenting this word correctly

Several approaches have been proposed to address the domain adaption problem for CWS. One major family proposed to compose boundary features by fitting the relevance of consecutive characters using Accessor Variety (AV) (Feng et al., 2004a,b), or Chi-square Statistics (Chi2) (Chang and Han, 2010). Combining the boundary features with other hand-crafted features, these methods were shown to achieve better performance on OOV words.

Inspired by these models, we propose a novel BLSTM-based neural network model which incorporates a global recurrent structure designed to model boundary features dynamically. This structure can learn to utilize the target domain corpus and extract the correlation or irrelevance between characters, which is a reminiscence of the discrete boundary features such as Accessor Variety (AV).

The contributions of this paper are two folds:

- First, we propose a global recurrent structure and incorporate it in the BLSTM-based neural network model for CWS. The structure can capture correlations between characters, and thus is especially efficient for segmenting OOV words and enhancing the performance of CWS on non-news domains.

- We obtain competitive results comparing to the best reported in the literature on the SIGHAN Bakeoff 2010 data, which is a benchmark dataset for cross-domain CWS.

2 BLSTM Architecture for Chinese Word Segmentation

We regard Chinese word segmentation task as a character-based sequence labeling problem, by labeling each character a tag from {S, B, E, M}. These tags indicate the position of the character in the segmented word. B, E, M represents *Begin, Middle, End* of a multi-character segmentation respectively, while S represents a single-character segmentation.

Figure 1 illustrates the general BLSTM architecture for Chinese word segmentation.

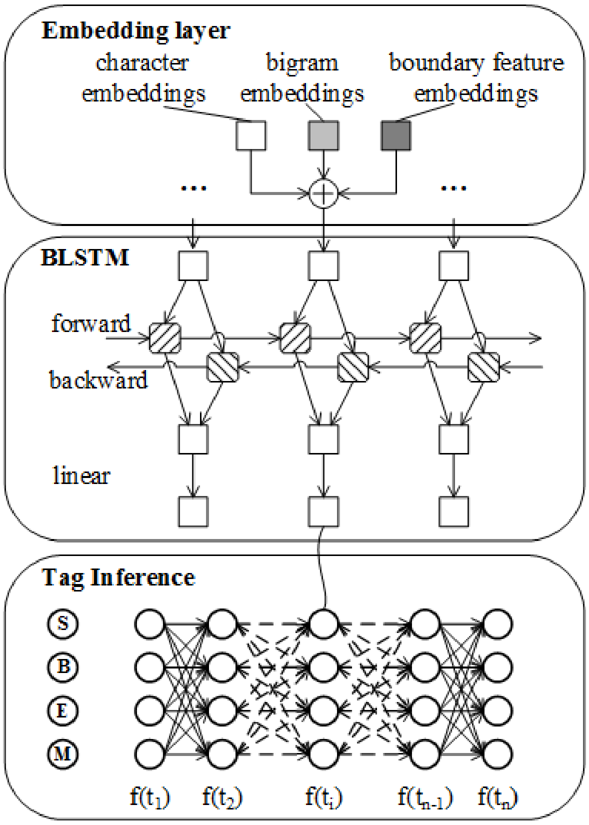


Figure 1: General architecture for Chinese word segmentation.

2.1 Embeddings

The outputs of the embedding layer is a concatenation of three parts: character embeddings, bigram embeddings and boundary feature embeddings.

We adopt the the local window approach which assumes that the tag of a character largely depends on its neighboring characters. For each character c_i in a given input sentence $c_{[1:n]}$, the context characters $c_{[i-w/2:i+w/2]}$ and their corresponding bigrams are chosen to be fed into the networks, where w is the context window size. As most CWS methods do, we will set $w = 5$ in our experiments.

Given a character set V of size $|V|$, each character $c \in V$ will be mapped into a d-dimensional embedding space as $Emb_c(c) \in \mathbb{R}^d$ by a lookup table $\mathbf{M_c} \in \mathbb{R}^{d \times |V|}$. Similarly, each bigram $b \in \{c_1c_2|c_1 \in V, c_2 \in V\}$ will be mapped into a d-dimensional embedding space as $Emb_b(b) \in \mathbb{R}^d$ by a lookup table $\mathbf{M_b} \in \mathbb{R}^{d \times |V| \times |V|}$.

The boundary feature embeddings are hidden vectors computed from the current bigrams and the whole bigarm history, which will be explained in detail in Section 3.

Three kinds of embeddings of the context characters $c_{[i-2:i+2]}$ and their corresponding bigrams are then concatenated into a single vector $x_i \in \mathbb{R}^{H_1}$, where $H_1 = 5d+4d+4d_{bf}$. d_{bf} is the number of hidden units output by the boundary feature embeddings. Then, this vector x_i is fed into the BLSTM layer.

2.2 Bidirectional LSTM Network

Following the embedding layer is an one-layer BLSTM network (Graves and Schmidhuber, 2005). By combining hidden states from two separate LSTM layers, it can incorporate long periods of contextual information from both directions. The LSTM cell is implemented as follows (Graves et al., 2013):

$$\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)\\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)\\
c_t &= f_tc_{t-1} + i_ttanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)\\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)\\
h_t &= o_ttanh(c_t)
\end{aligned} \tag{1}$$

where σ is the logistic sigmoid function, and i, f, o and c are the input gate, forget gate, output gate and the cell respectively, all of which are the same dimension as the hidden output h. The subscripts of the weight matrix describe the meaning as the name suggests. For instance, W_{xi} is the input gate weight matrix for input x.

The outputs of the BLSTM layer are the concatenation of a forward hidden sequence $\overrightarrow{h}$ and a backward hidden sequence $\overleftarrow{h}$ which will be fed to the decoding layer that contains a linear transformation with no non-linear function:

$$f(t_i|c_{[i-w/2:i+w/2]}) = W_d(\overrightarrow{h_i} \oplus \overleftarrow{h_i}) + b_d \tag{2}$$

where $W_d \in \mathbb{R}^{|T|\times H_2}$, $b_d \in \mathbb{R}^{|T|}$. H_2 is the number of hidden units of the outputs for the BLSTM layer. $f(t_i|c_{[i-w/2:i+w/2]}) \in \mathbb{R}^{|T|}$ is the score vector for each possible tag. Here in Chinese word segmentation, we set $T = \{S, B, E, M\}$.

2.3 Tag Inference

To model the correlations between tags in neighborhoods and jointly decode the best chain of tags for a given sentence, a transition score A_{ij} is introduced to measure the probability of jumping from tag $i \in T$ to tag $j \in T$ (Collobert et al., 2011). For an input sentence $c_{[1:n]}$ with a tag sequence $t_{[1:n]}$, a sentence-level score can be formulated as follows:

$$s(c_{[1:n]}, t_{[1:n]}, \theta) = \sum_{i=1}^{n}(A_{t_{i-1}t_i} + f_\theta(t_i|c_{[i-2:i+2]})) \tag{3}$$

where $f_\theta(t_i|c_{[i-2:i+2]})$ indicates the score output for the ith tag computed by the neural network described above with parameters θ.

3 Global Recurrent Structure

Chinese word segmentation is essentially a task of resolving the relevance of consecutive characters. Lacking knowledge of such relevance, recognizing out-of-domain words has been the bottleneck of domain adaption in CWS. However, Boundary features such as Accessor Variety (AV) (Feng et al., 2004a,b), Mutual Information (Sun and Xu, 2011) and Chi-square Statistics (Chi2) (Chang and Han, 2010) are features designed to fit such relevance. A significant advantage of boundary features is that they can compute the correlation of characters from a large scale corpora, annotated or not, to boost the OOV word recognition performance. As a result, they are especially effective for cross-domain CWS.

In this paper, we propose 5 novel global recurrent structures to generate embeddings that mimic the boundary features for further computing, which needs minimal pre-processing and feature engineering. The structures are designed to capture the intuition that nearby sentences in a single-domain corpus often share certain words. Thus the correlation of characters within or across certain words can be learned, and those involving OOV words notably enhance domain adaption for CWS.

GRS-1 The basic structure(GRS-1) is illustrated in Figure 2. It looks like LSTM-2 (Chen et al., 2015b) when incorporated into the BLSTM model. However, the difference is that common recurrent networks will reset the hidden states every time they process a new sentence in NLP problems while the hidden states in our structure are never reset.

$$h_{k+1,0}, c_{k+1,0} = h_{k,n_k}, c_{k,n_k} \tag{4}$$

where $h_{k,i}$ and $c_{k,i}$ are the hidden state and cell vector of the kth sentence at the ith step, n_k is the

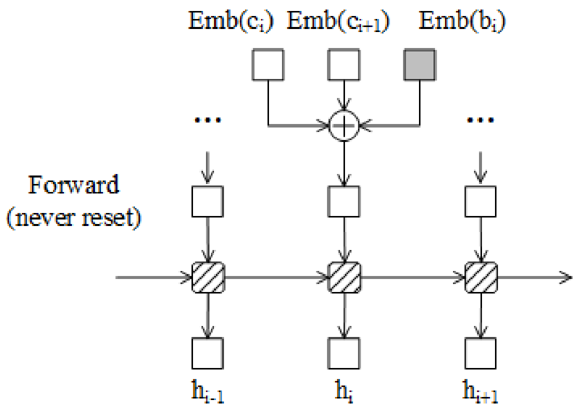


Figure 2: Global recurrent structure.

length of the kth sentence. For simplicity, in the following part we will ignore the subscript k and always indicate the current sentence.

As a result of such warm start mechanism, our structure can to some extent record the history information in recent sentences. And some information may last long in the structure.

Here we choose the LSTM cell as it can learn to keep relatively long term memory. We follow the equations (1) to implement it and directly take h_i as the boundary feature embeddings for the bigram $b_i = c_i c_{i+1}$ in the basic structure, where the input is the concatenation of embeddings of a bigram and its corresponding characters $Emb_b(b_i) \oplus Emb_c(c_i) \oplus Emb_c(c_{i+1})$.

$$Emb_{bf}(b_i) = h_i \tag{5}$$

where h_i is the output of the recurrent network at the ith step.

We also propose four more variants of the structure that are shown in Figure 3.

GRS-2 To better fit the boundary features, we add a full-connection hidden layer following the recurrent network. The boundary feature embeddings are calculated as follows:

$$Emb_{bf}(b_i) = \sigma(W_{bf} h_i + b_{bf}) \tag{6}$$

where σ is the logistic sigmoid function.

GRS-3 Considering the hidden states are noisy and contains much information of other words, we want the hidden values more relevant to the current bigram, so a gate is introduced to the structure. The boundary feature embeddings are calculated as follows:

$$\begin{aligned} E_i &= Emb_c(c_i) \oplus Emb_c(c_{i+1}) \oplus Emb_b(b_i) \\ g(b_i) &= \sigma(W_g E_i + b_g) \\ Emb_{bf}(b_i) &= g(b_i) h_i \end{aligned} \tag{7}$$

where $\oplus$ is the symbol for concatenation.

GRS-4 GRS-4 is a combination version of GRS-2 and GRS-3 by adding a full-connection hidden layer following the gated output.

GRS-5 GRS-5 is a more complicated version which tries to mimic the **Accessor Variety(AV)** criterion. AV criterion is a feature describing the number of distinct characters that precede or succeed a certain string s. For simplicity, we only focus on strings with $length = 2$, in other words, bigrams. Therefore, we substitute the input of GRS-4 with a bigarm and its preceding character to fit its left AV and similarly with a bigram and its succeeding character to fit its right AV. At last, we simply concatenate the two embeddings as the final boundary feature embeddings (Actually they are trigram boundary feature embeddings):

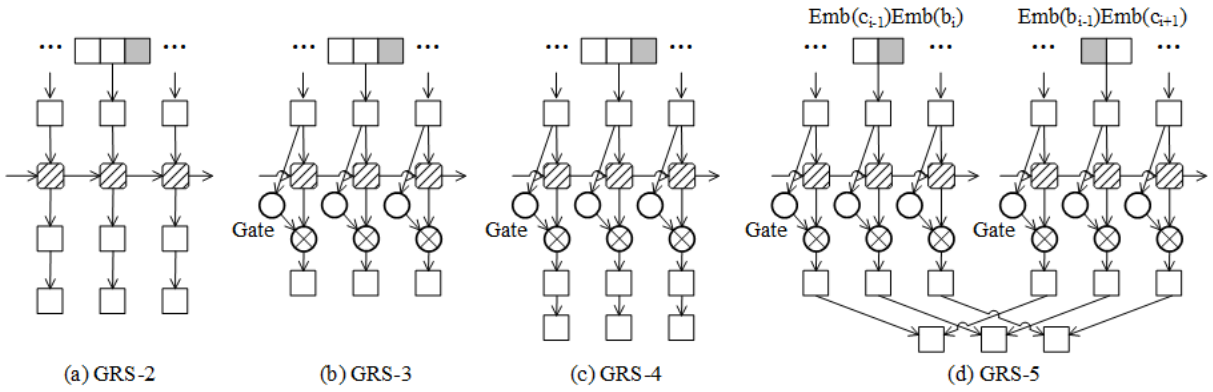
$$\begin{aligned} E_i^L &= Emb_c(c_{i-1}) \oplus Emb_b(b_i) \\ E_i^R &= Emb_c(c_{i+1}) \oplus Emb_b(b_{i-1}) \\ g^L(tri_i) &= \sigma(W_g^L E_i^L + b_g^L) \\ g^R(tri_i) &= \sigma(W_g^R E_i^R + b_g^R) \\ Emb_{bf}^L(tri_i) &= \sigma(W_{bf}^L g^L(tri_i) h_i^L + b_{bf}^L) \\ Emb_{bf}^R(tri_i) &= \sigma(W_{bf}^R g^R(tri_i) h_i^R + b_{bf}^R) \\ Emb_{bf}(tri_i) &= Emb_{bf}^L(tri_i) \oplus Emb_{bf}^R(tri_i) \end{aligned} \tag{8}$$

where $tri_i = c_{i-1} c_i c_{i+1}$ and other values have the same meanings as above.

4 Training

Instead of using the Max-Margin criterion (Taskar et al., 2005) adopted by previous neural network models for CWS (Zheng et al., 2013; Pei et al., 2014; Chen et al., 2015a,b), we try to directly maximize the log-probability of the correct tag sequence following Lample et al. (2016):

$$\begin{aligned} log(p(y|X)) &= s(X,y) - log(\sum_{\tilde{y} \in Y_X} e^{s(X,\tilde{y})}) \\ &= s(X,y) - \operatorname*{logadd}_{\tilde{y} \in Y_X} s(X,\tilde{y}) \end{aligned} \tag{9}$$



(a) GRS-2 (b) GRS-3 (c) GRS-4 (d) GRS-5

Figure 3: Four variants of the global recurrent structure.

where Y_X represents all possible tag sequences for a sentence X. While decoding, we predict the output sequence which obtains the maximum score as follows:

$$y^* = \underset{\tilde{y} \in Y_X}{\operatorname{argmax}}\, s(X, \tilde{y}) \tag{10}$$

The optimal sequence can be computed using dynamic programming. We use Adam (Kingma and Ba, 2014) to maximize the objective function.

5 Experiments

5.1 Experimental Setup

Data. We use the PKU corpus drawn from news domain for the source-domain training. The PKU dataset is provided by SIGHAN Bakeoff 2005 (Emerson, 2005). We regard the random 90% sentences of the training data as training set and the rest 10% sentences as development set. We also use the test part of the PKU dataset to measure the in-domain segmentation ability of our models. Following Liu et al. (2014)'s settings, our domain adaption experiments are performed on the four testing sets from the SIGHAN Bakeoff 2010 (Zhao and Liu, 2010) whose domains cover finance, computer, medicine and literature. In addition, we manually annotate six more corpora from non-news domains as testing sets, including finance, medicine, geology, agriculture, material and weather domains, which are extracted from abstracts of papers in **CNKI**[1]. These data are annotated following the guideline proposed by Yu et al. (2001). The OOV rate of these data are relatively high because they are more academic. Statistics of the training and testing data are shown in the Table 1.

All datasets are pre-processed by replacing the Chinese idioms and the continuous English letters and digits with a unique token.

Embeddings. We use *word2vec*[2] to pre-train character embeddings on the training corpus. The bigram embeddings are initialized with the average of the corresponding two characters' embeddings.

Discrete Boundary Features. The discrete boundary features which will be used in Section 5.3 are extracted from the datasets mentioned above and the Chinese Gigaword corpus[3], following methods in Sun and Xu (2011)'s paper.

Hyper-parameters. The hyper-parameters are tuned according to the experimental results. The detailed values are shown in Table 2.

Character & bigram embedding size	100
Boundary feature embedding size	100
Hidden unit number(cell in GRS)	300
Hidden unit number(cell in BLSTM)	300
Batch size	10
Early stop	5
Initial learning rate	0.02
Dropout rate on input layer	0.2
Regularization	10^{-4}

Table 2: Settings of the hyper-parameters.

5.2 Model Selection

We evaluate the baseline BLSTM model and our five proposed structures with the parameter settings in Table 2 on the PKU test data and six do-

[1] http://www.cnki.net/

[2] http://word2vec.googlecode.com/

[3] https://catalog.ldc.upenn.edu/LDC2003T05

Dataset	Train	Test	Test-Bakeoff2010			
	PKU		Finance	Computer	Medicine	Literature
# of Sent.	19056	1945	561	1330	1309	671
# of Words	1109947	104372	33035	35319	31499	35735
OOV Rate		0.0575	0.0874	0.1522	0.1102	0.0694

Dataset	Test-CNKI					
	Finance	Medicine	Geology	Agriculture	Material	Weather
# of Sent.	100	100	100	100	100	100
# of Words	27549	37803	29251	28780	26778	27228
OOV Rate	0.0437	0.2247	0.1910	0.1689	0.2224	0.1449

Table 1: Statistics of datasets used in this paper.

mains from the **CNKI** dataset. The results are shown in Table 3. The BLSTM+GRS-4 model with a gate and an additional full-connection hidden layer achieves the best performances among all domains. Surprisingly, the most delicate structure GRS-5 seems to be of no help to the CWS task.

To examine whether OOV recognition can benefit from GRS, we also look into the IV and OOV recalls of the PKU dataset respectively. Table 4 and Table 5 show that the proposed GRS can effectively improve the segmentation performance on OOV words, which empirically proves its domain adaption ability. BSLTM-2, similar to LSTM-2 (Chen et al., 2015b), is an architecture comprised of two stacking bidirectional LSTM hidden layers. GRS-4 is short for BLSTM+GRS-4 model.

Methods	IV Recall	OOV Recall
BLSTM	**97.12**	83.01
BLSTM-2	96.89	82.59
GRS-4	96.91	**83.78**

Table 4: IV and OOV recalls on the PKU development data.

Methods	IV Recall	OOV Recall
BLSTM	**96.35**	82.67
BLSTM-2	96.11	82.01
GRS-4	96.25	**83.96**

Table 5: IV and OOV recalls on the PKU test data.

5.3 Final Results

In this section, We compare our BLSTM+GRS-4 model with previous state-of-the-art methods.

Experimental results on the four test domains from SIGHAN Bakeoff 2010 are shown in Table 6. We also attempt to integrate discrete boundary features into the models. In our experiments, we choose the **Accessor Variety(AV)** (Feng et al., 2004a,b) which is a feature widely used in traditional Chinese word segmentation. Our F-scores and OOV recalls are competitive to those reported by Liu et al. (2014) and Jiang et al. (2013). However, following Liu et al. (2014)'s setting, we choose the PKU dataset as the training corpus while Jiang et al. (2013)'s model is trained on a different corpus. The results are not directly comparable. The results prove the incredible effectiveness of the global recurrent structure on OOV recognition and overall segmentation, comparable to the BLSTM model that directly incorporates discrete AV features. Adding discrete AV features into our model seem not to be a notable improvement, which also confirms that our model already has certain domain adaption ability.

Models	PKU	MSRA
(Zheng et al., 2013)	92.8	93.9
(Pei et al., 2014)	95.2	97.2
(Chen et al., 2015a)	96.4	97.6
(Chen et al., 2015b)	**96.5**	97.4
(Chen et al., 2015a)*	94.5	95.4
(Chen et al., 2015b)*	94.8	95.6
(Cai and Zhao, 2016)	95.5	96.5
(Zhang et al., 2016)	95.7	**97.7**
BLSTM	95.9	97.0
This work	95.9	97.1

Table 7: Comparison of our model with previous neural models on the PKU and MSRA datasets. Results with * are from runs on their released implementation (Cai and Zhao, 2016).

We compare the in-domain experimental results on the PKU and MSRA datasets with previ-

Dataset	Baseline(F%)	GRS-1(F%)	GRS-2(F%)	GRS-3(F%)	GRS-4(F%)	GRS-5(F%)
PKU	95.91	95.17	95.90	95.35	**95.92**	94.81
Out-of-Domain						
Finance	96.87	**97.15**	96.78	96.68	**97.15**	96.09
Medicine	85.01	86.24	85.98	85.83	**87.13**	85.97
Geology	87.59	88.52	88.90	88.38	**89.22**	86.90
Agriculture	89.51	90.54	91.16	90.90	**91.18**	90.08
Material	87.29	88.84	89.04	88.28	**89.62**	87.79
Weather	90.62	92.21	92.70	92.21	**93.21**	91.15

Table 3: Experimental results of the baseline BLSTM model and our proposed structures on the PKU test data and six domains from the **CNKI** dataset.

Method	Finance		Computer		Medicine		Literature		Avg-F	Avg-Roov
	F	Roov	F	Roov	F	Roov	F	Roov		
BLSTM	94.70	86.02	92.17	81.84	91.34	73.51	92.51	73.80	92.68	78.79
BLSTM+AV	95.77	90.91	93.57	82.82	92.50	83.12	93.79	**84.60**	93.91	**85.36**
GRS-4	**95.81**	**91.21**	**93.99**	83.81	92.26	**83.27**	**94.33**	81.30	**94.10**	84.90
GRS-4+AV	95.77	91.02	93.20	83.97	91.80	82.17	93.50	82.01	93.57	84.77
Liu2014	95.54	88.53	93.93	**87.53**	92.47	78.28	92.49	76.84	93.61	82.80
Jiang2013	93.16		91.19		**93.34**		93.53		92.80	

Table 6: Experimental results of the baseline BLSTM model, best-performance BLSTM+GRS-4 model, models with discrete AV features and models proposed by others on the SIGHAN Bakeoff2010 data.

ous neural models, which is shown in Table 7. The baseline BLSTM model with no modification or augmentation can achieve comparative results while the GRS does little help to the in-domain Chinese word segmentation task.

5.4 Error Analysis

We collect and analyze the errors on the Medicine corpus from Sighan Bakeoff 2010 in light of the fact that the results are the worst among the four domains. We calculate accuracies of individual OOV words, where accuracies are simply treated as 0 or 1 for further counting, and categorize them according to their frequencies in the testing corpus. Statistics are shown in Figure 4. From the trendlines we can infer that in our proposed GRS more occurrences yield higher accuracy while common BLSTM models can rarely benefit from this. That conforms to the intuition of our model that can utilize correlation information of testing corpora. Our model thereupon performs better with the increase of the size of testing corpus as long as the OOV words appear more.

Although the trendline of our model is promising, there are some OOV words that occurs frequently but are wrongly segmented. Some examples are listed in Table 8. Errors involving “肾脏”(kidney) and “维生素C”(vitamin C) are typical examples of the Combination Ambiguity, where there are some words containing “肾脏” such as “肾脏病学”(nephrology). Likewise, “维生素”(vitamin) is a frequent word that confuses our model. “甲型H1N1流感”(influenza A(H1N1)) reveals another severe problem that most CWS systems confront when processing the mix of Chinese characters and digits, punctuations or letters from other languages. The commonly used methods by treating consecutive digits or letters as one indeed boost the performance on corpora where most characters are Chinese. However, with the increase of characters other than Chinese, it is becoming a problem that should be reconsidered carefully.

OOV Word	English	Correct	Total
肾脏	kidney	15	39
甲型H1N1流感	influenza A (H1N1)	0	30
维生素C	vitamin C	2	23

Table 8: Some examples of wrongly segmented OOV words with high frequency.

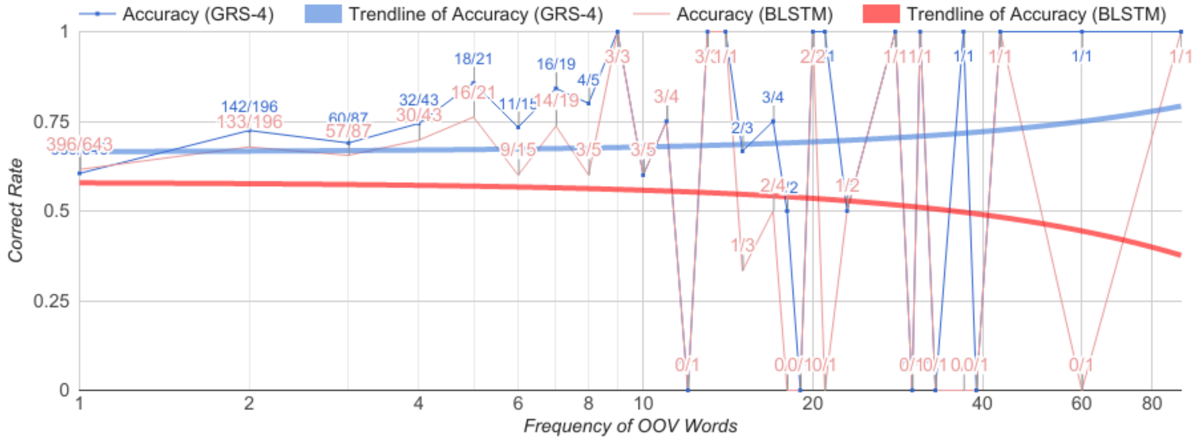


Figure 4: OOV word recognition accuracies on the Medicine corpus.

6 Related Work

Word segmentation has been pursued with considerable efforts in the Chinese NLP community. One mainstream method is regarding word segmentation task as a sequence labeling problem (Xue, 2003; Peng et al., 2004). Recently, researchers have tended to explore neural network based approaches (Collobert et al., 2011; Zheng et al., 2013; Qi et al., 2014) to reduce efforts of feature engineering. Pei et al. (2014) used a neural tensor model to capture the complicated interactions between tags and context characters. Experiments in his paper also show that bigram embeddings are of great benefit. To incorporate complicated combinations and long-term dependency information of the context characters, gated recursive model (Chen et al., 2015a) and LSTM model (Chen et al., 2015b) were used respectively. Moreover, Xu and Sun (2016) proposed a dependency-based gated recursive model which merges the benefits of the two models above. Coincidentally, Cai and Zhao (2016) and Zhang et al. (2016) both addressed the problem of lacking word-based features that previous neural CWS models have. Cai and Zhao (2016) proposed a novel gated combination neural network which thoroughly eliminates context windows and can utilize complete segmentation history. Zhang et al. (2016) proposed a transition-based neural model which replaces manually designed discrete features with neural features.

Domain adaption for Chinese word segmentation has been widely exploited before neural CWS models are proposed. Jiang et al. (2013) utilized the web text(160K Wikipedia) to improves segmentation accuracies on several domains. Zhang et al. (2014) studied type-supervised domain adaptation for Chinese segmentation by making use of domain-specific tag dictionaries and only unlabeled target domain data. Liu et al. (2014) proposed a variant CRF model to leverage both fully and partially annotated data transformed from different sources of free annotations consistently.

Some researches which focus on making use of unlabeled data for word segmentation also do help to domain adaption. Zhao and Kit (2008) and Zhang et al. (2013a) improved segmentation performance by mutual information between characters, collected from large unlabeled data. Li and Sun (2009) used punctuation information in a large raw corpus to learn a segmentation model, and achieve better recognition of OOV words. Sun and Xu (2011) explored several statistical features derived from both unlabeled data to help improve character-based word segmentation. Zhang et al. (2013b) proposed a semi-supervised approach that dynamically extracts representations of label distributions from both in-domain corpora and out-of-domain corpora.

7 Conclusion and Perspectives

In this paper, we propose a novel global recurrent structure to model dynamic boundary features and incorporate it in the BLSTM-based neural network model for Chinese Word Segmentation. The structure can capture correlations between characters, and thus is especially effective for segmenting OOV words and enhancing the performance of CWS on non-news domains.

The proposed global recurrent structure is not limited to the Chinese word segmentation task. It can be easily adapted to other sequence labeling problems that may benefit from the history information carried in the structure.

Although the structure is effective in this task, it's admittedly hard to train a stable model. As our future work, we would like to try some pretraining methods to handle this problem. And we plan to apply our method to other natural language processing tasks, such as Name Entity Recognition (NER). Also, the hybrid model is a great idea to try and we will do it later.

Acknowledgments

Our work is supported by National Natural Science Foundation of China (No.61370117 & No.61433015). The corresponding author of this paper is Houfeng Wang.

References

Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22(1):39–71.

Deng Cai and Hai Zhao. 2016. Neural word segmentation learning for chinese. *CoRR* abs/1606.04300.

Baobao Chang and Dongxu Han. 2010. Enhancing domain portability of chinese segmentation model using chi-square statistics and bootstrapping. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 789–798.

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)*. Association for Computational Linguistics, Beijing, China, pages 1744–1753.

Xinchi Chen, Xipeng Qiu, Chenxi Zhu, Shiyu Wu, and Xuanjing Huang. 2015b. Sentence modeling with gated recursive neural network. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 793–798.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research* 999888:2493–2537.

Thomas Emerson. 2005. The second international chinese word segmentation bakeoff. In *Proceedings of the Second SIGHAN Workshop on Chinese Language Processing*. pages 123–133.

Haodi Feng, Kang Chen, Xiaotie Deng, and Weimin Zheng. 2004a. Accessor variety criteria for chinese word extraction. *Computational Linguistics, Volume 30, Number 1, March 2004* .

Haodi Feng, Kang Chen, Chunyu Kit, and Xiaotie Deng. 2004b. Unsupervised segmentation of chinese corpus using accessor variety. In *Natural Language Processing - IJCNLP 2004, First International JointConference, Hainan Island, China, March 22-24, 2004, Revised Selected Papers*. pages 694–703.

Alex Graves, Abdel rahman Mohamed, and Geoffrey E. Hinton. 2013. Speech recognition with deep recurrent neural networks. *CoRR* abs/1303.5778.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks* 18:602–610.

Wenbin Jiang, Meng Sun, Yajuan Lü, Yating Yang, and Qun Liu. 2013. Discriminative learning with natural annotations: Word segmentation as a case study. In *ACL*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*. ICML '01, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *HLT-NAACL*.

Zhongguo Li and Maosong Sun. 2009. Punctuation as implicit annotations for chinese word segmentation. *Computational Linguistics* 35:505–512.

Yijia Liu, Yue Zhang, Wanxiang Che, Ting Liu, and Fan Wu. 2014. Domain adaptation for crf-based chinese word segmentation using free annotations. In *EMNLP*.

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)*. Association for Computational Linguistics, Baltimore, Maryland, pages 293–303.

Fuchun Peng, Fangfang Feng, and Andrew Mccallum. 2004. Chinese Segmentation and New Word Detection using Conditional Random Fields. In *Proceedings of COLING*. pages 562–571.

Yanjun Qi, Sujatha G. Das, Ronan Collobert, and Jason Weston. 2014. Deep learning for character-based information extraction. In *ECIR*.

Weiwei Sun and Jia Xu. 2011. Enhancing chinese word segmentation using unlabeled data. In *EMNLP*.

Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. 2005. Learning structured prediction models: a large margin approach. In *ICML*.

Jingjing Xu and Xu Sun. 2016. Dependency-based gated recursive neural network for chinese word segmentation. In *ACL*.

Nianwen Xue. 2003. Chinese Word Segmentation as Character Tagging. *Computational Linguistics* 8(1):29–48.

Shiwen Yu, Jianming Lu, Xuefeng Zhu, Huiming Duan, Shiyong Kang, Honglin Sun, Hui Wang, Qiang Zhao, and Weidong Zhan. 2001. Processing norms of modern chinese corpus. *Technical report* .

Longkai Zhang, Li Li, Zhengyan He, Houfeng Wang, and Ni Sun. 2013a. Improving chinese word segmentation on micro-blog using rich punctuations. In *ACL*.

Longkai Zhang, Houfeng Wang, Xu Sun, and Mairgup Mansur. 2013b. Exploring representations from unlabeled data with co-training for chinese word segmentation. In *EMNLP*.

Meishan Zhang, Yue Zhang, Wanxiang Che, and Ting Liu. 2014. Type-supervised domain adaptation for joint segmentation and pos-tagging. In *EACL*.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Transition-based neural word segmentation. In *A-CL*.

Hai Zhao and Chunyu Kit. 2008. An empirical comparison of goodness measures for unsupervised chinese word segmentation with a unified framework. In *IJCNLP*.

Hongmei Zhao and Qiu Liu. 2010. The cips-sighan clp2010 chinese word segmentation backoff.

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*. Association for Computational Linguistics, Seattle, Washington, USA, pages 647–657.