LM Enhanced BiRNN-CRF for Joint Chinese Word Segmentation and POS Tagging

Jianhu Zhang[1], Gongshen Liu[1](✉), Jie Zhou[1], Cheng Zhou[2],
and Huanrong Sun[2]

[1] School of Electric Information and Electronic Engineering,
Shanghai Jiaotong University, Shanghai, China
{zhangjianhu3290,lgshen,sanny02}@sjtu.edu.cn
[2] SJTU-Shanghai Songheng Content Analysis Joint Lab, Shanghai, China
{zhoucheng,sunhuanrong}@021.com

Abstract. Word segmentation and part-of-speech tagging are two preliminary but fundamental components of Chinese natural language processing. With the upsurge of deep learning, end-to-end models are built without handcrafted features. In this work, we model Chinese word segmentation and part-of-speech tagging jointly on the basis of state-of-the-art BiRNN-CRF architecture. LSTM is adopted as the basic recurrent unit. Apart from utilizing pre-trained character embeddings and trigram features, we incorporate neural language model and conduct multi-task training. Highway layers are applied to tackle the discordance issue of the naive co-training. Experimental results on CTB5, CTB7, and PPD datasets show the effectiveness of the proposed method.

Keywords: Chinese word segmentation · POS tagging · LSTM
Language model

1 Introduction

Word segmentation and part-of-speech(POS) tagging are two preliminary but fundamental components of Chinese natural language processing(NLP). Ng and Low [18] demonstrate that word segmentation and POS tagging could be modeled as a sequence labeling problem, in which target labels are the combinations of segmentation boundaries and POS tags, namely the joint S&T. Based on traditional handcrafted features, researchers have done a lot of remarkable work [8,10,30,33].

With the rapid development of the artificial neural network and deep learning, many neural models are applied to the joint S&T, reducing the efforts of feature engineering and boosting the performance. Currently, character-based BiRNN-CRF model achieves state-of-the-art performance on the Chinese joint S&T [2,32], in which bidirectional recurrent neural network(BiRNN) is the main

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11109, pp. 105–116, 2018.
https://doi.org/10.1007/978-3-319-99501-4_9

backbone for sequence tagging, and conditional random fields(CRF) [11] on top of BiRNN is used to gain the optimal tag sequence over the entire sentence. Moreover, LSTM [7] and GRU [3] are often applied to capture long-term relationship. Unlike previous pipeline methods, in which segmentation is put at the very first stage and then followed by POS tagging, the joint S&T model does not suffer from error propagation problem and word segmentation could make the most of information extracted from POS tagging to alleviate the ambiguity problem. Therefore, the joint S&T model outperforms the pipeline model significantly [2,24,33].

Recently, most researchers have focused on enriching the features of character embeddings [2,14,24,27,31]. However, given the complexity of neural network(NN) models and limited resources of the labeled corpus, it could be insufficient to train complicated models with annotations alone. Actually, there is a lot of semantic and syntactic knowledge that could be extracted from raw texts. Consequently, some semi-supervised and multi-task methods are proposed to improve sequence labeling performance [15,22,23].

In this paper, we extend the spirit of semi-supervising to the Chinese joint S&T and go a few steps further. Basic BiRNN-CRF framework for sequence labeling is adopted. We pre-train the Chinese character embeddings on large raw texts with GloVe [21]. In addition, we utilize n-gram embeddings to enrich the character features. More importantly, we propose to argument the current Chinese joint S&T architecture with a neural language model. The intuition is that it may be difficult for RNN to learn the proper representation as the hidden state considering the large number of parameters. Thus, it could be beneficial to add an extra but direct objective for the learning. The neural language model is simple and requires no additional data or annotation. It could be used as a training method and brings no extra computing cost during decoding.

To sum up, our contributions in this work are as follows:

1. Apply BiRNN-CRF model to the Chinese joint S&T and achieve state-of-the-art performance.
2. Propose to argument the current Chinese joint S&T architecture with a neural language model, which helps RNN to learn the proper hidden state but requires no additional data or annotation.
3. Conduct extensive experiments on three different datasets. Experimental results show that our best model outperforms previous state-of-the-art models.

2 LM Enhanced BiRNN-CRF Model

2.1 Overview of the Proposed Model

As visualized in Fig. 1, the proposed model is an adaptation of BiRNN-CRF enhanced by a neural language model. For a Chinese sentence $X = (c_1, \ldots, c_n)$, where c_i is the i-th character, all characters are represented as vectors and

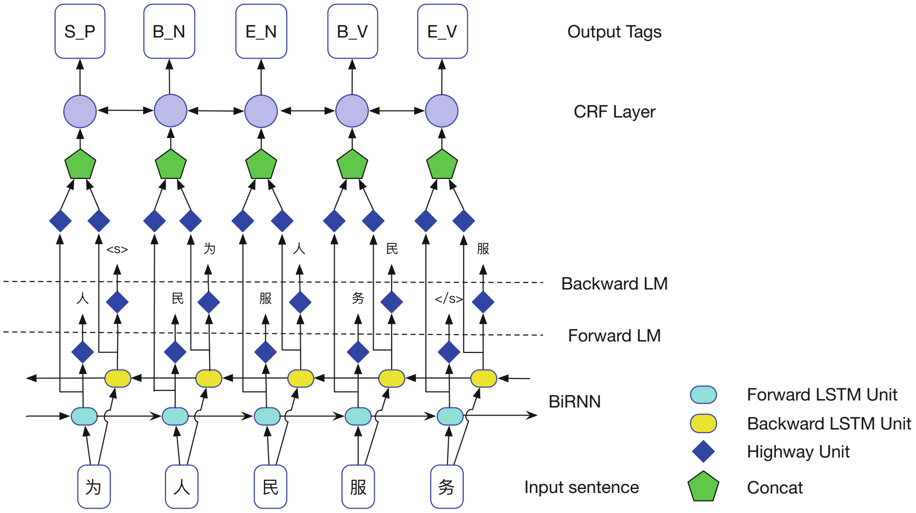

Fig. 1. LM Enhanced BiRNN-CRF Model Architecture

fed into BiRNN. LSTM is adopted as the basic recurrent unit. We employ the dropout [25] strategy to the output of BiRNN and then feed it into the first-order CRF layer. The sentence-level optimal tag sequence is predicted at the end. As for the tagging scheme, following the work of Kruengkrai et al. [10], we employ `B`, `I`, `E`, `S` as the word boundary tags, which represent a character at the beginning, inside, end of a word or a single character word respectively. In addition, in order to make the most of raw texts, output of BiRNN is also used as the input of a neural language model, which predicts the previous or next character. Highway layers are placed between output of BiRNN and target tasks to map the hidden state to different semantic space.

2.2 Character Representations

Character representations is of wide interest in Chinese NLP. Sun et al. [27] propose to enhance Chinese character embeddings with radical information, Li et al. [13] develop two component-enhanced Chinese character embedding models and their bigram extensions. Shao et al. [24] propose three different approaches to effectively represent Chinese characters, namely the concatenated n-gram, radicals and orthographical features as well as the pre-trained character embeddings. In order to keep the model as simple as possible, only pre-trained character embeddings and n-gram embeddings are employed in this paper.

n-gram Embeddings. Although RNN is good at extracting contextual features from context-free character representations, many researchers have shown that traditional n-gram embeddings is beneficial for many NLP tasks [5,12,29].

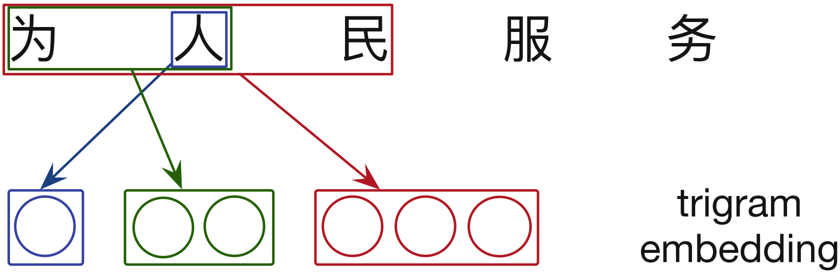

Fig. 2. Trigram embedding of a Chinese character in a given context

n-gram embeddings employed in this paper is demonstrated in Fig. 2. In this example, trigram embeddings of the pivot character 人 in the given context is the concatenation of the context-free vector representation of 人 itself along with the bigram 为人 as well as the trigram 为人民 .

Pre-trained Character Embeddings. The context-free vector representations of single characters used above could be replaced by pre-trained character embeddings trained from raw texts. In this paper, we pre-train character embeddings on Wikipedia[1] using GloVe [21]. This kind of pre-trained character embeddings is used to initialize the very bottom input of our neural networks. For those characters that are not in the embedding vocabulary, we randomly initialize them.

2.3 Neural Language Model

Although the character embeddings described above could carry a lot of general language knowledge, it is not task-specific, thus may contain a considerable irrelevant portion and maybe not optimal for the Chinese joint S&T. Moreover, due to the large number of parameters, it may be difficult for BiRNN to learn the proper hidden state. In order to address these problems, we propose to incorporate a language model with the joint S&T model and conduct multi-task learning.

As shown in Fig. 1, the language model and the S&T model share the same BiRNN, which fits the setting of multi-task learning and transfer learning. However, these two tasks are apparently not strongly related, which means directly using the output from the recurrent neural network layer could hurt the performance of the joint S&T model. Thus, highway layer [26] is applied to further transform the hidden state of RNN into different semantic space for different objectives. The highway layer can be illustrated as follows:

$$H(\mathbf{X}) = g(W_H\mathbf{X} + b_H) \odot T(\mathbf{X}) + \mathbf{X} \odot C(\mathbf{X}), \tag{1}$$

[1] https://dumps.wikimedia.org/.

where operator $\odot$ indicates element-wise product, $g(\cdot)$ is a certain type of nonlinear transformation. $T(\cdot)$ represents the transform gate and $C(\cdot)$ is the carry gate. In our experiments, $C(\cdot) = 1 - T(\cdot)$ is adopted. Transform gate $T(\cdot)$ could be formalized as:

$$T(\mathbf{X}) = \sigma(W_T \times \mathbf{X} + b_T) \tag{2}$$

where W_T and b_T are both trainable parameters.

In order to extract language knowledge from both directions, we adopt two language model, namely the forward language model(i.e., from left to right) and the backward language model(i.e., from right to left), which makes predictions for the next and previous character respectively. The forward language model is defined as

$$P_f(c_1, \ldots, c_n) = \prod_{i=1}^{N} P_f(c_i | c_1, \ldots, c_{i-1}) \tag{3}$$

where $P_f(c_i | c_1, \ldots, c_{i-1})$ is computed by the neural network with the following formula

$$P_f(c_i | c_0, \ldots, c_{i-1}) = \frac{\exp(w_{c_i}^T f_{i-1})}{\sum_{\hat{c}_j} \exp(w_{\hat{c}_j}^T f_{i-1})} \tag{4}$$

where w_{c_i} is the weight vector for predicting character c_i, f_{i-1} is output of corresponding highway unit. Consequently, the average negative log probability of the target words is applied as the object function of the forward language model:

$$\mathcal{J}_{F-LM} = -\frac{1}{N} \sum_i \log P_f(c_i) \tag{5}$$

Accordingly, the backward language model is defined as

$$P_b(c_1, \ldots, c_n) = \prod_{i=1}^{N} P_b(c_i | c_{i+1}, \ldots, c_N) \tag{6}$$

where $P_b(c_i | c_{i+1}, \ldots, c_N) = \frac{\exp(w_{c_i}^T b_i)}{\sum_{\hat{c}_j} \exp(w_{\hat{c}_j}^T b_i)}$. And the loss function is calculated as

$$\mathcal{J}_{B-LM} = -\frac{1}{N} \sum_i \log P_b(c_i) \tag{7}$$

With Eqs. 5 and 7, the overall objective function of our language model could be written as

$$\mathcal{J}_{LM} = \mathcal{J}_{F-LM} + \mathcal{J}_{B-LM} \tag{8}$$

2.4 CRF

Since we model the joint S&T as a sequence labeling task, it is beneficial to consider the correlations between labels in the neighborhood and jointly decode the optimal label sequence for a given input sentence. Therefore, we build a

CRF layer upon the BiRNN to decode each label jointly instead of independently. Formally, we use $z = (z_1, \ldots, z_n)$ to represent the BiRNN output(transformed by highway layer and concatenated with both directions) for an input sentence $x = (c_1, \ldots, c_n)$. $y(z) = (y_1, \ldots, y_n)$ is the corresponding label sequence for z. The probabilistic model of CRF describes the conditional probability of generating the whole label sequence y given z. Similar to Ma and Hovy [16], we define this probability with the following form:

$$p(y|z; W, b) = \frac{\prod_{i=1}^{n} \psi_i(y_{i-1}, y_i, z)}{\sum_{y' \in Y(z)} \prod_{i=1}^{n} \psi_i(y'_{i-1}, y'_i, z)} \tag{9}$$

where $\psi_i(y', y, z) = \exp(W^T_{y',y} z_i + b_{y',y})$ are potential functions, and $W^T_{y',y}$ and $b_{y',y}$ are the weight vector and bias corresponding to label pair (y', y), respectively.

In the training phase, the maximum conditional likelihood is applied, i.e., minimize the negative log-likelihood as follows:

$$\mathcal{J}_{CRF} = -\sum_i \log p(y_i|z_i) \tag{10}$$

In the testing or decoding phase, our target is to search for the optimal label sequence y^* with the maximum conditional probability:

$$y^* = \underset{y \in Y(z)}{\mathrm{argmax}}\, p(y|z; W, b) \tag{11}$$

In this paper, we employ first-order CRF layer; the Viterbi algorithm can solve the training and decoding in an efficient way.

2.5 Multi-task Learning

With Eqs. 8 and 10, our multi-task loss function could be written as

$$\mathcal{J} = \mathcal{J}_{CRF} + \lambda \mathcal{J}_{LM} \tag{12}$$

where λ is a weight parameter to make a balance between the language model and the joint S&T model.

3 Experiments

We conduct our experiments on three different datasets: Chinese Treebank 5.1(CTB5), Chinese Treebank 7.0(CTB7), as well as PKU's People's Daily (PPD). Table 1 summarizes the brief statistics of these corpora in terms of sentence number and word number. CTB5 is split according to [8], CTB7 is split according to [30], and PPD datasets are from the Fourth International Chinese Language Processing Bakeoff [9]. We apply the standard F1-score to evaluate both the word segmentation and the joint S&T performance.

Table 1. Datasets' statistics

Datasets	Splits	Num of sentences	Num of words
CTB5	Train	18,086	494 k
	Dev	350	6.8 k
	Test	348	8.0 k
CTB7	Train	31,088	718 k
	Dev	10,036	237 k
	Test	10,291	245 k
PPD	Train	66,691	1.1 M
	Test	6,424	160 K

Table 2. Hyper-parameters.

Char. embedding size	64
LSTM state size	200
LSTM depth	1
Highway depth	1
Optimiser	Adagrad
Initial learning rate	0.1
Decay rate	0.05
Gradient clipping	5.0
Dropout rate	0.5
Batch size	20

3.1 Implementation

We implement our model based on the TensorFlow library [1]. Bucket strategy is adopted in our word, namely grouping the sentences of similar length into the same bucket and pad them to the equivalent length accordingly. Each bucket has their own computational graph.

The hyper-parameters adopted in this paper is shown in Table 2. We use the same hyper-parameters for all the experiments on different datasets without further fine-tuning. The weights of the neural networks are initialized as Glorot and Bengio [6]. We use the error back propagation algorithm to train the network. Mini-batch stochastic gradient descent with momentum is employed for optimization [4]. We apply dropout in our model, and the dropout rate is fixed at 0.5. Gradient clipping [19] of 5.0 is used for model stability.

3.2 Component Effects

In order to find out the effects of each component of our neural network architecture, we run some ablation experiments, and the results are shown in Table 3.

Table 3. Effects of each component on test sets of CTB5, CTB7, PDD datasets.

Models	CTB5			CTB7			PPD		
	P	R	F	P	R	F	P	R	F
BiLSTM-CRF	91.23	91.08	91.15	89.44	88.46	88.95	89.12	89.21	89.16
+ pre-trained embeddings	92.43	92.03	92.23	89.74	89.16	89.45	89.52	89.74	89.63
+ trigram embeddings	93.16	92.88	93.02	90.32	90.14	90.23	90.11	90.03	90.07
+ language model	93.04	92.92	92.98	90.82	90.04	90.43	89.93	89.89	89.91
+ highway layer	94.83	94.25	94.54	91.65	90.38	91.01	91.76	90.39	91.07

Table 4. Comparisons with previous models on test sets in word segmentation and joint S&T F1-scores.

Models	CTB5		CTB7		PPD	
	Seg	Seg&Tag	Seg	Seg&Tag	Seg	Seg&Tag
ZPar	97.69	93.79	94.59	89.71	94.62	89.94
Shap et al. [24]	97.98	94.06	95.37	90.54	94.95	90.76
Ours	98.72	94.88	95.74	91.24	95.45	91.32

The base model is BiRNN-CRF, in which character embeddings are randomly initialized. Pre-trained embeddings, n-gram embeddings, neural language model, and highway layer are incrementally incorporated into the base model.

Pre-trained Embeddings and Trigram Embeddings. From Table 3, we could learn that employing the pre-trained embeddings as initial vector representations for characters achieves improvements on all three datasets. Intuitively, the pre-trained character embeddings give a more reasonable initialization for NN. Further improvement is obtained by adding trigram embeddings, which reveals rich local information could be encoded in n-gram with a statistical co-occurrence way.

Language Model and Highway Layer. The third and fourth row of Table 3 elucidate the effects of the neural language model and the highway layer. As the third row shown, incorporating language model but without highway layer yields a little or no performance improvement. We conjecture the reason for this is that the joint S&T model and the language model are not strongly related so that they could not get benefits easily from each other. However, by employing highway layer, we get notable improvements. From these results, we could see that the language model is able to capture task-specific language knowledge, and the highway layer plays an important role in mediating the joint S&T model and the language model. In general, we incorporate all the essential components: pre-trained embeddings, trigram embeddings, neural language model, and highway layer into the basic BiRNN-CRF as our final model.

Table 5. Comparisons with [2] on test sets in joint S&T Precision, Recall and F1-scores.

Models	CTB5			CTB7			PPD		
	P	R	F	P	R	F	P	R	F
Chen et al. [2]	92.88	93.49	93.19	84.40	86.25	85.31	90.27	90.05	90.16
Ours	94.28	95.48	94.88	85.51	89.04	87.24	92.03	90.62	91.32

3.3 Performance Comparison

In this section, we focus on comparisons between the proposed model and previous state-of-the-art models. Ensemble decoding of three models trained independently is employed to get the best performance. Experimental results are shown in Table 4.

We first compare our model with ZPar [34], which is one of the most prevalent joint tagger using structured perceptron and beam decoding. We retrain a ZPar model and reproduce the performance reported in the original paper on CTB5. Then we train and evaluate ZPar on CTB7 and PPD respectively. As shown in Table 4, our model outperforms ZPar on all the three datasets. Theoretically, ZPar utilizes discrete local information at both character and word levels and employs structured perceptron for global optimization [32], while we employ BiRNN to model complex features and capture long-term dependencies. The CRF layer is applied on top of BiRNN for sentence level optimization. Moreover, we propose to incorporate a neural language model to conduct co-training to extract task-specific language knowledge from the raw text. We further employ highway layer to unify the joint S&T model and the language model. In contrast to the traditional methods, our model need no feature engineering or data preprocessing and benefits from large raw texts.

Recently, Shao et al. [24] propose several vector representations of Chinese characters to improve the joint S&T performance. We retrain their model on CTB5 using the best setting reported in their paper and reproduce their evaluation scores. We also retrain and evaluate their model on PDD and CTB7 for comparison. As shown in Table 4, besides without any feature engineering our model still outperforms theirs.

In Table 5, we compare our model with Chen et al. [2] in terms of the joint S&T F1 score. Chen et al. [2] introduce the convolution layer into the Chinese joint S&T and use different filters to model the complex compositional features of Chinese characters. We take the performance scores from their paper directly. Notably, they split the CTB7 dataset in a different way to estimate the model robustness. Following their setting, we also test our model on web blogs and train it on the rest of dataset. As shown in Table 5, our model outperforms theirs by a relatively large margin.

4 Related Work

As two of the most fundamental tasks in Chinese NLP, word segmentation and POS tagging have been studied for decades. Traditional methods like CRFs, HMMs, and maximum entropy classifiers [17,20,28] focused on feature engineering to create powerful handcrafted features for each specific task, which means that it is difficult to apply them to new tasks or domains. Recently, with the upsurge of deep learning, we have gotten rid of feature engineering but build end-to-end models. Meanwhile, CRF layer has also been demonstrated to be effective in capturing the dependency among labels and is widely used as the tag inference layer for sequence labeling tasks. Our model is built upon the success of BiRNN-CRF model and is further extended to better capture the character information for the Chinese S&T with a co-training neural language model.

Recently, Shao et al. [24] propose a character-based joint S&T model for Chinese using BiRNN-CRF. They mainly focus on enrich the character embeddings such as extracting radicals and orthographical features with convolutional neural networks(CNN), whereas we propose to enhance the current BiRNN-CRF model with a neural language model in a multi-task learning way. Peters et al. [22] propose to enrich word embeddings with pre-trained language model and obtain notable performance improvement. But this method requires expensive resources and time. Liu et al. [15] and Rei [23] both propose to empower sequence labeling with semi-supervised language model. We extend this spirit and modified the model architecture to make it applicable for the Chinese S&T.

5 Conclusion

In this paper, we model the Chinese S&T jointly as a fully character-based sequence labeling task. BiRNN-CRF is adopted and LSTM is used as the basic recurrent unit. In order to effectively extract language knowledge from the unstructured corpus, in addition to utilizing pre-trained character embeddings and trigram embeddings, we propose to incorporate neural language model and conduct multi-task training to help RNN learn the proper hidden state. Highway layers are applied to overcome the discordance issue of the naive co-training. Our model is extensively evaluated on CTB5, CTB7, and PPD datasets. The experimental results on the test sets show that the proposed model outperforms ZPar and other state-of-the-art models.

Acknowledgement. This research work has been funded by the National Natural Science Foundation of China (Grant No.61772337, U1736207 and 61472248), the SJTU-Shanghai Songheng Content Analysis Joint Lab, and program of Shanghai Technology Research Leader (Grant No.16XD1424400).

References

1. Abadi, M. et al.: Tensorflow: a system for large-scale machine learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, OSDI 2016, pp. 265–283. USENIX Association, Berkeley (2016)
2. Chen, X., Qiu, X., Huang, X.: A long dependency aware deep architecture for joint Chinese word segmentation and POS tagging. CoRR abs/1611.05384 (2016)
3. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR abs/1412.3555 (2014)
4. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. **12**, 2121–2159 (2011)
5. Durme, B.V., Rastogi, P., Poliak, A., Martin, M.P.: Efficient, compositional, order-sensitive n-gram embeddings. In: EACL (2017)
6. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Teh, Y.W., Titterington, M. (eds.) Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research, PMLR, Chia Laguna Resort, Sardinia, Italy, vol. 9, pp. 249–256, 13–15 May 2010
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735
8. Jiang, W., Huang, L., Liu, Q., Lü, Y.: A cascaded linear model for joint chinese word segmentation and part-of-speech tagging. In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (2008)
9. Jin, G., Chen, X.: The fourth international Chinese language processing bakeoff: Chinese word segmentation, named entity recognition and Chinese POS tagging. In: IJCNLP (2008)
10. Kruengkrai, C., Uchimoto, K., Kazama, J., Wang, Y., Torisawa, K., Isahara, H.: Joint Chinese word segmentation and POS tagging using an error-driven word-character hybrid model. IEICE Trans. Inf. Syst. **E92**(D12), 2298–2305 (2009). https://doi.org/10.1587/transinf.E92.D.2298
11. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning, ICML 2001, pp. 282–289. Morgan Kaufmann Publishers Inc., San Francisco (2001)
12. Li, B., Liu, T., Zhao, Z., Wang, P., Du, X.: Neural bag-of-ngrams. In: AAAI, pp. 3067–3074 (2017)
13. Li, Y., Li, W., Sun, F., Li, S.: Component-enhanced Chinese character embeddings. CoRR abs/1508.06669 (2015)
14. Ling, W., Luís, T., Marujo, L., Astudillo, R.F., Amir, S., Dyer, C., Black, A.W., Trancoso, I.: Finding function in form: Compositional character models for open vocabulary word representation. CoRR abs/1508.02096 (2015)
15. Liu, L., Shang, J., Xu, F.F., Ren, X., Gui, H., Peng, J., Han, J.: Empower sequence labeling with task-aware neural language model. CoRR abs/1709.04109 (2017)
16. Ma, X., Hovy, E.H.: End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. CoRR abs/1603.01354 (2016)
17. McCallum, A., Freitag, D., Pereira, F.C.N.: Maximum entropy markov models for information extraction and segmentation. In: Proceedings of the Seventeenth International Conference on Machine Learning, ICML 2000, pp. 591–598. Morgan Kaufmann Publishers Inc., San Francisco (2000)

18. Ng, H.T., Low, J.K.: Chinese part-of-speech tagging: one-at-a-time or all-at-once? word-based or character-based? In: Lin, D., Wu, D. (eds.) Proceedings of EMNLP 2004, pp. 277–284. Association for Computational Linguistics, Barcelona, July 2004
19. Pascanu, R., Mikolov, T., Bengio, Y.: Understanding the exploding gradient problem. CoRR abs/1211.5063 (2012)
20. Peng, F., Feng, F., McCallum, A.: Chinese segmentation and new word detection using conditional random fields. In: Proceedings of the 20th International Conference on Computational Linguistics, COLING 2004. Association for Computational Linguistics, Stroudsburg (2004). https://doi.org/10.3115/1220355.1220436
21. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
22. Peters, M.E., Ammar, W., Bhagavatula, C., Power, R.: Semi-supervised sequence tagging with bidirectional language models. CoRR abs/1705.00108 (2017)
23. Rei, M.: Semi-supervised multitask learning for sequence labeling. CoRR abs/1704.07156 (2017)
24. Shao, Y., Hardmeier, C., Tiedemann, J., Nivre, J.: Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF. CoRR abs/1704.01314 (2017)
25. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**, 1929–1958 (2014)
26. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. CoRR abs/1505.00387 (2015)
27. Sun, Y., Lin, L., Tang, D., Yang, N., Ji, Z., Wang, X.: Radical-enhanced Chinese character embedding. CoRR abs/1404.4714 (2014)
28. Uchiumi, K., Tsukahara, H., Mochihashi, D.: Inducing word and part-of-speech with Pitman-Yor hidden semi-Markov models. In: ACL (2015)
29. Vajjala, S., Banerjee, S.: A study of n-gram and embedding representations for native language identification. In: Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, pp. 240–248 (2017)
30. Wang, Y., Kazama, J., Tsuruoka, Y., Chen, W., Zhang, Y., Torisawa, K.: Improving Chinese word segmentation and POS tagging with semi-supervised methods using large auto-analyzed data. In: IJCNLP (2011)
31. Wieting, J., Bansal, M., Gimpel, K., Livescu, K.: Charagram: Embedding words and sentences via character n-grams. CoRR abs/1607.02789 (2016)
32. Yang, Z., Salakhutdinov, R., Cohen, W.W.: Transfer learning for sequence tagging with hierarchical recurrent networks. CoRR abs/1703.06345 (2017)
33. Zhang, Y., Clark, S.: Joint word segmentation and POS tagging using a single perceptron. In: Proceedings of ACL-08: HLT, pp. 888–896. Association for Computational Linguistics, Columbus, June 2008
34. Zhang, Y., Clark, S.: A fast decoder for joint word segmentation and POS-tagging using a single discriminative model. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, pp. 843–852. Association for Computational Linguistics, Stroudsburg (2010)