IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101201

A Transformer-based Neural Model for Chinese Word Segmentation and Part-of-Speech Tagging

Xinxin Li

School of Computer Science and Technology, Shandong University of Technology, Zibo, China

Abstract: Recently, deep learning methods have greatly improved the state-of-the-art in many natural language processing tasks. Previous work shows that the Transformer can capture long-distance relations between words in a sequence. In this paper, we propose a Transformer-based neural model for Chinese word segmentation and part-of-speech tagging. In the model, we present a word boundary-based character embedding method to overcome the character ambiguity problem. After the Transformer layer, BiLSTM-CRF layer is used to generate the best tagging results. Experiments on Chinese Treebank show that our model on Chinese word segmentation and part-of-speech tagging outperforms the baseline model and achieves state-of-the-art performance.

Keywords: Chinese Word Segmentation, POS Tagging, Transformer, Word Boundary-Based Character Embedding

I. INTRODUCTION

Chinese word segmentation (CWS) and part-of-speech (POS) tagging are important tasks in Chinese natural language processing (NLP). Traditionally, word segmentation is a preliminary step before performing part-of-speech tagging. To avoid the error propagation from word segmentation to part-of-speech tagging and improve the interaction between two tasks, these two tasks are commonly trained as a joint model, such as character-based method [1], word-based method [2], word-character hybrid method [3], sub-word method [4], and linear and lattice reranking methods [5], [6].

In recent years, neural network models have been shown to improve Chinese word segmentation and part-of-speech tagging. Zheng introduced neural networks for CWS and POS tagging and presented a perceptron-based method to train the model [7]. Shao proposed a bidirectional RNN-CRF architecture for Chinese word segmentation and POS tagging. It is a character-based model and utilizes rich contextual information and sub-character level features [8]. Zhao presented a lattice-LSTM and Convolutional Network, which can exploit multi-granularity of information, including characters, words, and subwords [9]. Tian introduced a neural network model with a two-way attention mechanism. It incorporated context features and the corresponding syntactics of each character in the sequence [10]. RNN, LSTM, and GRU networks have achieved good results in various natural language processing tasks [11], such as part-of-speech tagging [12], syntactic parsing [13], [14], semantic role labeling [15], and machine translation [16]. However, the inherent attribute of RNN hinders the training parallelization. Attention mechanism allows modelling the dependencies of input and output sequences without considering their distance in the sequence [17], [18]. The Transformer is a model that avoids recurrence. It completely relies on the attention mechanism to model the global dependency between input and output.

Besides the structures of neural network models for NLP tasks, pre-trained embedding is another important factor that is incorporated into neural network models. Chinese characters have ambiguity problems since one character might have different meanings in different words. To better utilize this information, we introduce word boundary-based character embedding, splitting each character into four different boundaries. We propose a Transform-based neural model for Chinese word segmentation and part-of-speech tagging using word boundary-based character embedding.

II. RELATED WORK

A. Chinese Word Segmentation and Part-of-Speech Tagging

Chinese word segmentation and part of speech tagging are the basic problems in Chinese natural language processing. Traditional methods viewed these two problems as two separate sequence labeling problems. Xue proposed a word-based model for Chinese word segmentation and used the maximum entropy model to train the model [19]. Peng et al. applied the conditional random field model to Chinese word segmentation and new word detection and introduced domain information and word combination information [20]. Andrew proposed a hybrid model based on a conditional random field and a semi-Markov conditional random field and applied it to Chinese word segmentation [21]. Zhang and Clark a proposed a word-based learning method for Chinese word segmentation using the structured perceptron method, and used the beam search method to obtain the best tagging sequence [22]. Tang et al. proposed a large margin method for Chinese word segmentation, which effectively improved the prediction performance [23]. Zhao and Kit

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 10, Issue 12, December 2021
DOI: 10.17148/IJARCCE.2021.101201

introduced a variety of unsupervised features into Chinese word segmentation and adopted a conditional random field model to effectively combine the information [24]. Jiang et al. proposed a method based on structured perceptron method to integrate the word information in Wikipedia [25].

Recently, joint learning methods have been used in Chinese word segmentation and part of speech tagging, so that part of speech information can also be used in word segmentation. Ng and Low proposed a label combination method that combines word boundary labels with part of speech labels, converts the multi-task sequence labeling problems into a single sequence labeling problem and then uses the maximum entropy model to learn [1]. Jiang et al. applied the minimum error learning method to this problem and introduced models such as the co-occurrence frequency model and word n-gram [6]. Zhang and Clark proposed a word-based structured perceptron method, introduced word and part of speech information into the model, and used the beam search method to obtain the best results [2]. Kruengkrai et al. adopted a word lattice and proposed error-driven learning strategy, which can predict known and unknown words [3]. Sun et al. proposed a subword-based method. The method uses a secondary structure, where the first level uses multiple word segmentation models to generate subword sequences, and the second level uses subword sequences to train the model [4]. Zheng et al. applied the neural network model to Chinese word segmentation and part of speech tagging and proposed a perceptron-based method to train the model [26]. Shao proposed a bidirectional RNN-CRF architecture that incorporated rich contextual information and sub-character level features [8]. Zhao presented a model based on lattice-LSTM and Convolutional Network, exploiting character, word, and subword information [9]. Tian proposed a two-way attention neural network model using context features and their corresponding syntactic information of characters in the sequence [10].

B. Chinese Character Embedding

The word embeddings are learned from distributional information of word contexts in large corpora, such as the skip-gram model, Glove [27]–[29]. However, this method has two limitations. First, it only takes linear contexts. To overcome the disadvantage, Levy and Goldberg improved the skip-gram model by incorporating dependency relations as linear contexts [28]. Wang introduced an approach that represented entities and words/phrases by jointly embedding knowledge graphs and a text corpus. Rothe presented AutoExtend method that extended word embeddings to synsets/lexemes embeddings, and be able to adapt on various resources [30]. The second drawback of the skip-gram model is that it cannot handle the polysemy problem well. Liu presented a latent topic models, which first chose the topics for each word in the text data, and then trained topical word embeddings on words and their topics [31]. For Chinese language, characters are more suitable than words in NLP tasks, and the polysemy problem is severer than in English. Shi proposed a radical embedding model that decomposed each character into 4 radicals and applied it on several NLP tasks [32].

To overcome the second disadvantage, this paper proposes a word boundary-based character embedding for Chinese sequence labeling problems. In Chinese, one character often doesn't have a particular meaning except that it is in a specific word or as a single character word. Characters have different meanings when they are concatenated into different words or sentences. For example, the character 家 ("home") in the word 作家 ("writer") and word 家庭 ("family") as prefix and suffix expresses different semantic meanings. Boundary information can be used to distinguish characters. To better utilize this information, we introduce word boundary-based character embedding, splitting each character into four components, each with different boundary tags.

III. THE MODEL

The method of sequence labeling is to label each character in the sequence with one of the given tags. For Chinese word segmentation, a character is assigned as one of four boundary tags (B, M, E, S), separately representing the beginning character, the middle character, the end character of a word, and a single character word. For word segmentation task, characters in the word 服务员 (waiter) are annotated as B, M, E separately. For Chinese word segmentation and POS tagging, the label of each character is the combination of boundary tag and POS tag, e.g., B-NN, M-NN, E-NN.

We proposed a Transformer-based neural network model for Chinese word segmentation and POS tagging. The architecture of the model is shown in Fig. 1. Different from utilizing traditional character embedding, our neural model uses word boundary-based character embedding. Then The model adopts a Transformer layer to capture long-distance relations among words in the sentence. The best tag sequence for the sentence can be obtained with a BiLSTM-CRF layer.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101201

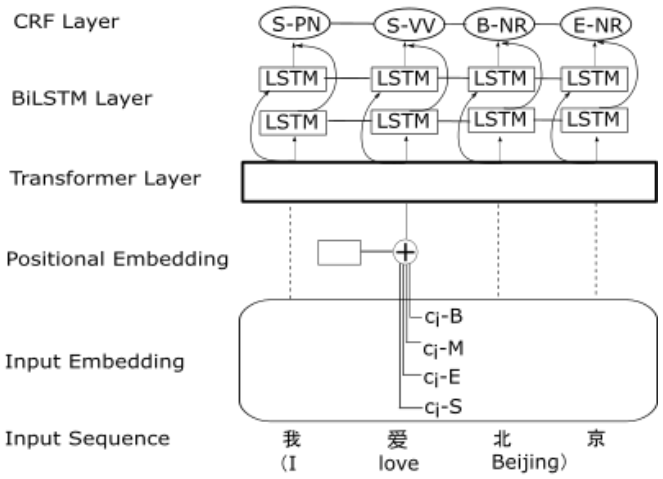

Fig. 1 The architecture of our model

A. Embedding Layer

The first layer of the network takes character sequences as the input, and transforms them into real-valued vectors. In common, the size of characters is fixed, and stored in a dictionary D. For each character $i \in D$, its corresponding real-valued embedding can be retrieved through $LT_c(.)$,

$$LT_c(i) = C_i$$

where $C_i \in R^{|S|}$ is the corresponding real-valued embedding for character i, and $|S|$ is the dimension of character embedding. C is the matrix denoting the vector representations of all characters, where the i^{th} row is the real-valued vector for character i. Then, a character sequence $[c]_1^n = (c_1, c_2, ..., c_n)$ can be transformed into vector sequence $LT_c([c]_1^n) = (C_1, C_2, ..., C_n)$. We learn character embeddings using GloVe method [29].

For Chinese, a word is the basic unit to represent syntactic and semantic meaning. Some characters have specific meanings as single-character words, but might change their meanings when combining into words, such as the character 老 ("older") in the word 老师 ("teacher"). Some characters might not have specific meanings, but can form meaning words, such as the character 幽 in the word 幽默 ("humor"). These characters usually need to be combined together to form a word to represent a specific meaning. We observe that characters in different positions within words tend to exhibit different syntactic and semantic properties. Based on this viewpoint, we propose a word boundary-based character embedding for our task. A character is split into four components according to the boundary tag. Our raw text is annotated by a state-of-the-art word segmenter and transformed into word boundary-based text. For example, the segmented sentence 骑/自行车/到/银行 (ride a bicycle to the bank) is annotated as "骑_S 自_B 行_M 车_E 到_S 银_B 行_E". Then the character 行 (can) in word 自行车 (bicycle) is distinguished from the word 银行 (bank).

For Chinese word segmentation and POS tagging, the embedding of each character c is then composed of its split four parts, simply the concatenation of new character embeddings $c_E = (c-B, c-M, c-E, c-S)$. To use the relative or absolute position for each character, we use positional embedding c_P and perform bitwise addition with c_E to obtain the final embedding.

B. Transformer Layer

The Transformer layer in our model contains N stacks of the same layer, and each layer has two sublayers. The first sublayer uses multi-head self-attention networks, and the second sublayer is a connected feedforward network. These two sublayers are followed by layer normalization and connected by residual connections.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101201

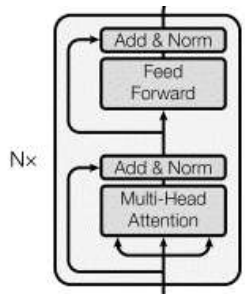

Fig. 2 Transformer Layer

The multi-head attention networks can learn from different perspectives on the input.

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)W^O,$$

$$where\, head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

Where $W_i^Q \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^K \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^V \in \mathbb{R}^{d_{model}\times d_v}$ and $W^O \in \mathbb{R}^{hd_v\times d_{model}}$. The dimension of the output is d_{model}. There are h heads, where the dimension $d_k = d_v = d_{model}/h$.

The attention is Scaled Dot-Product Attention. Its input includes query q and key k with dimension d_k, and value v with dimension d_v.

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

C. BiLSTM Layer

The output of the Transformer is used as the input of the BiLSTM layer. A typical LSTM unit is shown in Fig. 3. It contains an input gate i_t, a forget gate f_t, an output gate o_t, a memory cell C_t and a hidden unit h_t. It is calculated as

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

$$h_t = o_t \times \tanh(C_t)$$

For the BiLSTM layer, the hidden unit is the concatenation of forward LSTM and backward LSTM, and it is used as input for the next layer.

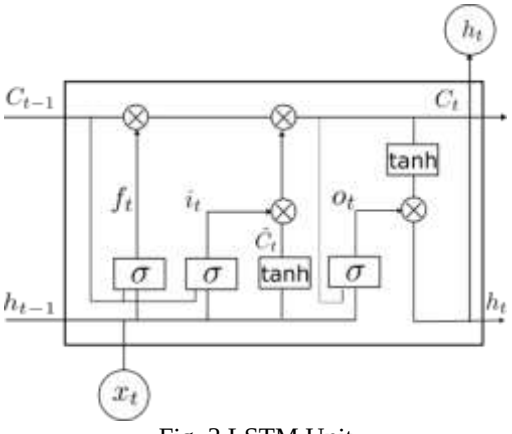

Fig. 3 LSTM Unit

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101201

D. CRF Layer

The CRF layer is used to predict final tagging sequence. Given input sequence X, the probability of the output sequence y is denoted as $p(y|x)$. $p(y|x)$ is a log-linear model,

$$p(y|x;w) = \frac{\exp(w \cdot \Phi(x,y))}{\sum_{y' \in Y^m} \exp(w \cdot \Phi(x,y'))}$$

Where $\Phi(x,y) \in \mathbb{R}^d$ is feature vector. To simplify the model complexity, we define the feature vector $\Phi(x,y)$ as

$$\Phi(x,y) = \sum_{j=1}^{m} \phi(x, j, y_{j-1}, y_j)$$

IV. EXPERIMENTS

A. Dataset and Experimental Setting

We perform our experiments on CTB 5.0 [33], where the distribution of the training, development, and test datasets is shown in table 1. The standard F-1 measure is used to evaluate the performance of word segmentation and overall segmentation and tagging. F-1 measure is the balance between precision P and recall R, defined as $F = 2PR/(P+R)$.

TABLE 1: The distribution of training, development and test dataset

Dataset	Training	Dev	Test
Chapter IDs	1-270, 400-931, 1001-1151	301-325	271-300
#sentences	18089	352	348
#words	493939	6821	8008
#POS tags	35		
OOV(word)	/	0.0811	0.0347
OOV(word&POS)	/	0.0874	0.0420

B. Experimental Results

To investigate the impact of the dimension of character embedding on the model, we perform experiments with different dimensions. The results are shown in Fig. 4. It's observed that with fixed training data, the performance improves with the increase of the dimension of character embedding. However, the training time of both the language model and our model increases along with the embedding dimension. Therefore, we take |S|=400 in the following experiments.

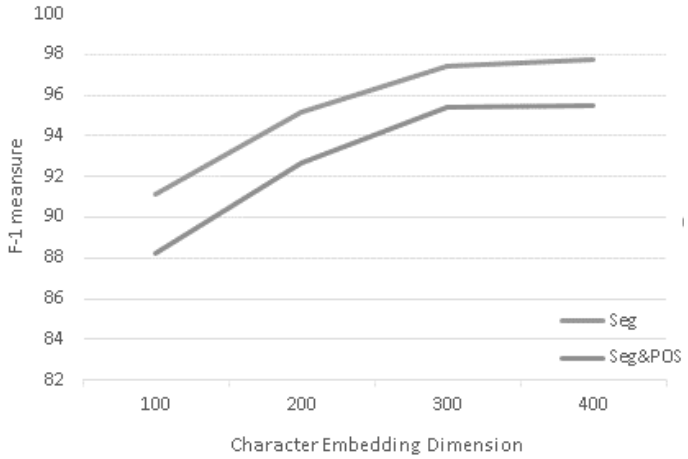

Fig. 4 Experimental results with different character embedding dimensions

We then study how well the word boundary-based character embedding enhances the baseline one. The character embedding is trained using Glove model. For word boundary-based character embedding, the raw data is first segmented by a state-of-the-art segmenter and transformed into word boundary-based text, as described in section 2.1. We use ICTCLAS tools in our experiments. The experimental results of our model on development data are shown in Table 2. We compare our model with/without word boundary-based character embedding (WBCE). It can be seen that

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101201

the model with word boundary-based character embedding outperforms the baseline model. It increases about 0.32% F-1 value on word segmentation and 0.19% F1 value on joint word segmentation and part-of-speech tagging.

TABLE 2: Experimental results of different methods on development data

Methods	Seg F1	Seg&POS F1
Wang [34]	96.28	93.16
Kruengkrai [3]	96.42	92.88
Shao [8]	97.42	94.58
Our Model-WBCR	97.34	95.23
Our Model	97.66	95.42

Table 3 shows the results of our model and other approaches. Our model performs better than Zheng's neural model, Jiang's word lattice model, Kruengkrai's word-character hybrid model, and Shao's BiRNN-CRF model. The performance of our system is still lower than Tian's model. The reason might be that their model uses rich contextual information such as n-gram features or sentential representations.

TABLE 3: Experimental results of different methods on test data

Methods	Seg F1	Seg&POS F1
Zheng [7]	95.23	91.82
Jiang [6]	97.85	93.41
Kruengkrai [8]	97.87	93.67
Shao [8]	97.87	95.23
Tian [10]	98.81	96.92
Our Model	98.12	96.25

V. CONCLUSION

We presented a neural model for Chinese word segmentation and part-of-speech tagging. The model used word boundary-based character embedding for each character and introduce Transformer encoder to capture long-distance relations between characters in the sequence. The final tagging sequence is predicted by the BiLSTM-CRF layer. Experimental results on CTB 5.0 show that our model with word boundary-based character embedding outperforms the baseline model, and achieves state-of-the-art performance.

REFERENCES

[1]. H. T. Ng and J. K. Low, “Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based?,” in Proceedings of EMNLP 2004, Barcelona, Spain, Jul. 2004, pp. 277–284.

[2]. Y. Zhang and S. Clark, “Joint Word Segmentation and POS Tagging Using a Single Perceptron,” in Proceedings of ACL-08: HLT, Columbus, Ohio, Jun. 2008, pp. 888–896.

[3]. C. Kruengkrai, K. Uchimoto, J. Kazama, Y. Wang, K. Torisawa, and H. Isahara, “An Error-Driven Word-Character Hybrid Model for Joint Chinese Word Segmentation and POS Tagging,” in Joint Conference of ACL and IJCNLP, Suntec, Singapore, 2009, pp. 513–521.

[4]. W. Sun, “A Stacked Sub-Word Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging,” in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, Jun. 2011, pp. 1385–1394.

[5]. W. Jiang, L. Huang, Q. Liu, and Y. Lü, “A Cascaded Linear Model for Joint Chinese Word Segmentation and Part-of-Speech Tagging,” in Proceedings of ACL-08: HLT, Columbus, Ohio, Jun. 2008, pp. 897–904.

[6]. W. Jiang, H. Mi, and Q. Liu, “Word Lattice Reranking for Chinese Word Segmentation and Part-of-Speech Tagging,” in Proceedings of the 22nd International Conference on Computational Linguistics, Manchester, UK, Aug. 2008, pp. 385–392.

[7]. X. Zheng, H. Chen, and T. Xu, “Deep Learning for Chinese Word Segmentation and POS Tagging,” in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, Washington, USA, Oct. 2013, pp. 647–657.

[8]. Y. Shao, C. Hardmeier, J. Tiedemann, and J. Nivre, “Character-based joint segmentation and POS tagging for Chinese using bidirectional RNN-CRF,” in Proceedings of the eighth international joint conference on natural language processing (volume 1: Long papers), Taipei, Taiwan, Nov. 2017, pp. 173–183.

[9]. L. Zhao, A. Zhang, Y. Liu, and H. Fei, “Encoding multi-granularity structural information for joint Chinese word segmentation and POS tagging,” Pattern Recognition Letters, vol. 138, pp. 163–169, 2020

[10]. Y. Tian et al., “Joint Chinese word segmentation and part-of-speech tagging via two-way attentions of auto-analyzed knowledge,” in Proceedings of the 58th annual meeting of the association for computational linguistics, Online, Jul. 2020, pp. 8286–8296. doi: 10.18653/v1/2020.acl-main.735.

[11]. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural Language Processing (Almost) from Scratch,” Journal of Machine Learning Research, vol. 12, no. 1, pp. 2493–2537, Nov. 2011.

[12]. M. Labeau, K. Löser, and A. Allauzen, “Non-lexical neural architecture for fine-grained POS Tagging,” in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, Sep. 2015, pp. 232–237.

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 10, Issue 12, December 2021

DOI: 10.17148/IJARCCE.2021.101201

[13]. G. Durrett and D. Klein, "Neural CRF Parsing," 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), Beijing, China, Jul. 2015, pp. 302–312.
[14]. M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy, and N. A. Smith, "Retrofitting Word Vectors to Semantic Lexicons," in Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, Jun. 2015, pp. 1606–1615.
[15]. J. Zhou and W. Xu, "End-to-end learning of semantic role labeling using recurrent neural networks," 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), Beijing, China, Jul. 2015, pp. 1127–1137.
[16]. L. Miculicich, D. Ram, N. Pappas, and J. Henderson, "Document-Level Neural Machine Translation with Hierarchical Attention Networks," in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, Oct. 2018, pp. 2947–2954. doi: 10.18653/v1/D18-1325.
[17]. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in 3rd international conference on learning representations, ICLR 2015, san diego, CA, USA, may 7-9, 2015, conference track proceedings, 2015.
[18]. A. Vaswani et al., "Attention is all you need," in Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, long beach, CA, USA, Dec. 2017, pp. 5998–6008.
[19]. N. Xue, "Chinese Word Segmentation as Character Tagging," Computational Linguistics and Chinese Language Processing, vol. 8, no. 1, pp. 29–48, 2003.
[20]. F. Peng, F. Feng, and A. McCallum, "Chinese Segmentation and New Word Detection using Conditional Random Fields," in Proceedings of the 20th international conference on Computational Linguistics, Geneva, Switzerland, 2004, pp. 562–568.
[21]. G. Andrew, "A Hybrid Markov/Semi-Markov Conditional Random Field for Sequence Segmentation," in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, Jul. 2006, pp. 465–472.
[22]. Y. Zhang and S. Clark, "Chinese Segmentation with a Word-Based Perceptron Algorithm," in Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Prague, Czech Republic, Jun. 2007, pp. 840–847.
[23]. B. Tang, X. Wang, and X. Wang, "Chunking with Max-Margin Markov Networks," in Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation, The University of the Philippines Visayas Cebu College, Cebu City, Philippines, Nov. 2008, pp. 474–480.
[24]. H. Zhao and C. Kit, "Integrating Unsupervised and Supervised Word Segmentation: The Role of Goodness Measures," Information Sciences, vol. 181, no. 1, pp. 163–183, 2011
[25]. W. Jiang, M. Sun, Y. Lü, Y. Yang, and Q. Liu, "Discriminative Learning with Natural Annotations: Word Segmentation as a Case Study," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, Aug. 2013, pp. 761–769.
[26]. M. Zhang, Y. Zhang, W. Che, and T. Liu, "Chinese Parsing Exploiting Characters," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, Aug. 2013, pp. 125–134.
[27]. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," in NIPS, Harrahs and Harveys, Lake Tahoe, Nevada, United States, 2013, pp. 1–9.
[28]. O. Levy and Y. Goldberg, "Dependency-Based Word Embeddings," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Baltimore, Maryland, Jun. 2014, pp. 302–308.
[29]. J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global Vectors for Word Representation," in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 2014, pp. 1532–1543.
[30]. S. Rothe and H. Schütze, "AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes," 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), Beijing, China, Jul. 2015, pp. 1793–1803.
[31]. Y. Liu, Z. Liu, T.-S. Chua, and M. Sun, "Topical Word Embeddings," in Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, Texas, USA, Jan. 2015, pp. 2418–2424.
[32]. X. Shi, J. Zhai, X. Yang, Z. Xie, and C. Liu, "Radical Embedding: Delving Deeper to Chinese Radicals," in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China, Jul. 2015, pp. 594–598.
[33]. N. Xue, "Annotating discourse connectives in the Chinese Treebank," in Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky, Ann Arbor, Michigan, 2005, pp. 84–91.
[34]. Y. Wang, J. Kazama, Y. Tsuruoka, W. Chen, Y. Zhang, and K. Torisawa, "Improving Chinese Word Segmentation and POS Tagging with Semi-supervised Methods Using Large Auto-Analyzed Data," in Proceedings of 5th International Joint Conference on Natural Language Processing, Chiang Mai, Thailand, Nov. 2011, pp. 309–317.

