Chinese Named Entity Recognition with Hidden State of Boundary

Anonymous ACL submission

Abstract

Chinese word boundaries cannot be directly displayed as Chinese is a sequence of characters. To attend words in sentences, inspired by span-based NER and boundary module in NER, the hidden states of current character come from its context in BiLSTM and are activated by sigmoid gate to represent boundaries. The boundaries are added into encode to get word-lever information of Chinese named entity. The values of boundaries are soft to show sentences structure obtained with labels. Experimental studies on four benchmark datasets and incorporated BERT for pre-training show our method gets the optimal recognition result in Chinese NER.

1 Introduction

Language equals speech plus structure, and without boundaries there is no structure. In contrast to English, Chinese is a sequence of characters. There is no separator between characters (Su et al., 2018; Li et al., 2014), so word boundaries cannot be directly displayed. However, word-level information is very important for natural language processing (Mao et al., 2008; Peng and Dredze, 2016b; Zhang and Yang, 2018). Different ways of defining words can lead to different word segmentation results. There are still some basic questions like "what is a word" and "a word is what" that are not answered. Research (Sproat et al., 1994) shows that even if one is a native Chinese speaker, the rate of agreement on words appearing in Chinese texts is only about 70%. Therefore, in a strict sense, automatic word segmentation is a problem that is not clearly defined.

Traditionally, for Chinese NER, Chinese Word segmentation(CWS) system is first performed (Yang et al., 2016; He and Sun, 2017b). However, the existing CWS output a large number of incorrect word segmentation results, which leads to unsatisfactory language processing, and do not perform well in recognizing long entities as they focus only on word-level information. In contrast to word-based partitioning methods, character-based partitioning methods (He and Wang, 2008; Liu et al., 2010; Li et al., 2014; Liu et al., 2019; Sui et al., 2019; Gui et al., 2019; Ding et al., 2019) have been empirically proven to be effective. A drawback of the purely character-based NER method is that the word information is not fully exploited. With this consideration, word lexicons are incorporated into the character-based NER model (Zhang and Yang, 2018; Peng et al., 2019; Li et al., 2020). However, they incorporate many wrong word lexicons without considering the whole sentences for segmentation. Moreover, as boundary detection and type prediction may cooperate with each other for the NER task, it is also important for the two subtasks to mutually reinforce each other by sharing their information.

To address the issue, we perform hidden state of boundary(HSB) to comparing with CWS system. HSB adopts a sequence to sequence model with RNN networks. Specifically, the model employs a bidirectional LSTM to model sequential dependencies of each character. The hidden states of current character come from its context and are activated by gate to represent boundaries. The boundaries are added into hidden states in forward to produce entity boundaries based on input sequence. Experimental results show our model outperforms on the performance. In summary, the main contributions of this paper include:

- We propose a simple but effective method for incorporating word boundaries into the character representations for Chinese NER.
- The proposed method is transferable to different sequence-labeling architectures and can be easily incorporated with pre-trained models like BERT (Devlin et al., 2018).

2 Background

2.1 Span-based NER

In the domain, NER is usually considered as a sequence labeling problem (Liu et al., 2018). Span-based NER identifies segments in a sentence and classifies each segment with a special label (e.g., PER, ORG or LOC). For boundary embedding, span representation is calculated by the concatenation of the start and end tokens' representations (Fu et al., 2021). To enumerate all possible text spans in a sentence, the concatenation of word representations of its startpoint and endpoint with a 20-dimensional embedding represent the span width (Li et al., 2021a) following previous work. Zhang et al. (2018) use adaptive co-attention network with LSTM structure. There are also inspiring tasks (Xu et al., 2021; Ma et al., 2022; Cao and Wang, 2022; Hong et al., 2022) about boundary representation that integrate better feature of language structure in model . More recently, pretrained language models such as ELMo and BERT have been adopted to further enhance the performance of NER.

2.2 Boundary Module in NER

The Boundary Module needs to provide not only distinct contextual boundary information but also segment information for the NER Module. Huang et al. (2015) utilize the BiLSTM as an encoder to learn the contextual representation of words, and then Conditional Random Fields (CRFs) is used as a decoder to label the words. It has achieved the state-of-the-art results on various datasets for the past many years. Inspired by the success of the BiLSTM-CRF architecture, many other state-of-the-art models have adopted such architecture. Li et al. (2021b) use another BiLSTM as encoder to extract distinct contextual boundary information. To add extra information to the input of the LSTM, they use the sum of the hidden states of current , previous and next words instead of word embedding. Specifically, the encoders obtain the distinct boundary hidden sequences and a sentinel vector is padded into the last positions of hidden sequences for the sentinel word inactive. Then, a unidirectional LSTM is used as a decoder to generate the decoded state at each time step. Li et al. (2021c) processes the starting boundary word in an entity to point to the corresponding ending boundary word. They train the starting boundary word to point to the corresponding ending boundary word, and the other words in the sentence to a sentinel word inactive. Boundary smoothing(Zhu and Li, 2022) applies the smoothing technique to entity boundaries, rather than labels. Smoothed boundary provides more continuous targets across spans, which are conceptually more compatible with the inductive bias of neural networks that prefers continuous solutions.

3 Method

3.1 LSTM-HSB structure

The character-based model uses an LSTM-CRF model on the character sequence $c_1, c_2, \ldots, c_m$. Each character c_t is represented using

$$\mathbf{x}_t^c = \mathbf{e}^c(c_t), \tag{1}$$

$\mathbf{e}^c$ denotes a character embedding lookup table.

The character representations are put into the sequence modeling layer, which models the dependency between characters. Generic architectures for this layer including the bidirectional long-short term memory network(BiLSTM), the Convolutional Neural Network(CNN) and the transformer(Vaswani et al., 2017). In this work, we implemented this layer with a single-layer Bi-LSTM. In our model, we use hidden states as boundaries.

Here, we precisely show the definition of the forward LSTM-HSB(Figure 1):

$$\begin{aligned}
\begin{bmatrix} i_t \\ f_t \\ o_t \\ \widetilde{C}_t \end{bmatrix} &= \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(W \begin{bmatrix} x_t^c \\ h_{t-1} \end{bmatrix} + b \right), \\
C_t &= \widetilde{C}_t \odot i_t + C_{t-1} \odot f_t, \\
h_t &= o_t \odot \tanh(C_t), \\
H_t &= h_t + \sigma(W^h h_t + b^h),
\end{aligned} \tag{2}$$

where σ is the element-wise sigmoid function and $\odot$ represents element-wise product. W, b, W^h and b^h are trainable parameters.

A bidirectional LSTM-HSB is applied to $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ to obtain $\overrightarrow{\mathbf{H}}_1^c, \overrightarrow{\mathbf{H}}_2^c, \ldots, \overrightarrow{\mathbf{H}}_m^c$ and $\overleftarrow{\mathbf{H}}_1^c, \overleftarrow{\mathbf{H}}_2^c, \ldots, \overleftarrow{\mathbf{H}}_m^c$ in the left-to-right and right-to-left directions, respectively, with two distinct sets of parameters. The backward LSTM-HSB shares the same definition as the forward LSTM-HSB yet model the sequence in a reverse order. The concatenated hidden states at the step of the forward and backward LSTM-HSBs forms the context-dependent representation. The hidden vector representation of each character is:

$$\mathbf{H}_t^c = [\overrightarrow{\mathbf{H}}_t^c; \overleftarrow{\mathbf{H}}_t^c]. \tag{3}$$

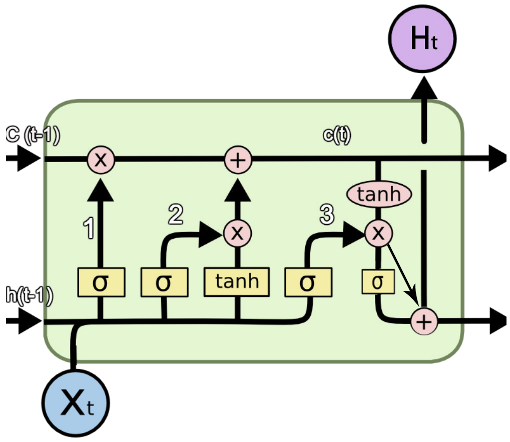

Figure 1: LSTM-HSB structure.

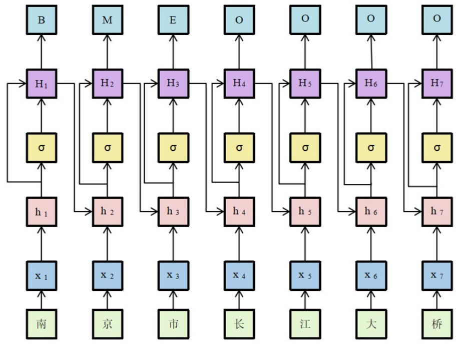

Figure 2: Character hidden boundary states model.

A standard CRF model is used on $\mathbf{H}^c_1, \mathbf{H}^c_2, \ldots, \mathbf{H}^c_m$ for sequence labelling.

By explicitly assigning probability to surrounding spans, the soft boundary of hidden states prevents the model from concentrating all probability mass on the scarce positive samples. This intuitively helps alleviate over-confidence. In addition, hard boundary presents noticeable sharpness between the classification targets of positive spans and surrounding ones, although they share similar contextualized representations. Soft boundary provides more continuous targets across spans, which are conceptually more compatible with the inductive bias of neural networks that prefers continuous solutions.

3.2 Hidden Boundary State Model of Chinese NER

We design the NER model (Figure 2) based on LSTM-HSB. Sequences of characters are pretrained to obtain character vectors. With Bi-LSTM, we get context hidden state. The hidden state is performed on the character vector to obtain the boundary between nearby characters. The boundary is activated and added with original hidden state to obtain the results. In addition, we show Table 1 for detail idea in model.

In Figure 2, the character sequences of ['南(South)', '京(Capital)', '市(City)', '长(Long)', '江(River)' , '大(Big)', '桥(Bridge)'],pretrained unigram and bigram embeddings, result in character vector groups $(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ respectively. With Bi-LSTM, we get context hidden states $(h_1, h_2, h_3, h_4, h_5, h_6, h_7)$. The hidden states are activated by sigmoid σ and added original hidden states to form boundary states $(H_1, H_2, H_3, H_4, H_5, H_6, H_7)$. The results are obtained with labels (B, M, E, O, O, O, O).

Table 1 indicates that boundaries come from the context hidden state. The value activated by sigmoid in the span of [0,1] is a soft way to locate boundary between nearby characters. Then is is added with original hidden states to obtain the results. Compering with input of boundary enhance(Table 2), our boundaries are dynamic and smoothed from hidden state activation without vocabulary.

4 Experiment

4.1 Setup

Datasets. The LSTM-HSB is evaluated on four Chinese NER datasets, including MSRA (Levow, 2006), OntoNotes (Weischedel et al., 2011), Resume NER (Zhang and Yang, 2018) and Weibo NER (Peng and Dredze, 2015; He and Sun, 2017a). Weibo NER is a social media domain dataset, which is drawn from Sina Weibo, while OntoNotes and MSRA datasets are in the news domain. Resume NER dataset consists of resumes of senior executives, which is annotated by (Zhang and Yang, 2018).

Evaluation. We use P, R and F1 in average to evaluate our performance on MSRA, OntoNotes and Resume datasets comparing with BERT-base and other methods. We used F1 in average to evaluate our performance on the NE, NM and Overall of Weibo dataset comparing with BERT-base and other methods.

Model settings. For LSTM-HSB model, we adopted similar settings as LSTM+CRF (`https://github.com/TVect/ChinNER/tree/master/models/lstm_crf`), LSTM+CRF+BERT (`https://github.com/`

$\mathbf{h}_t^c$	$\overrightarrow{\mathbf{h}}_t^c$	$\sigma(\mathbf{W}_f^h \overrightarrow{\mathbf{h}}_t^c + \mathbf{b}_f^h)$	$\overleftarrow{\mathbf{h}}_t^c$	$\sigma(\mathbf{W}_b^h \overleftarrow{\mathbf{h}}_t^c + \mathbf{b}_b^h)$	$\mathbf{H}_t^c$
南	南	0.1	南京市长江大桥	0.9	[南0.1;南0.9]
京	南京	0.8	京市长江大桥	0.8	[京0.8;京0.8]
市	南京市	0.9	市长江大桥	0.1	[市0.9;市0.1]
长	南京市长	0.8	长江大桥	0.8	[长0.8;长0.8]
江	南京市长江	0.6	江大桥	0.4	[江0.6;江0.4]
大	南京市长江大	0.2	大桥	0.8	[大0.2;大0.8]
桥	南京市长江大桥	0.8	桥	0.2	[桥0.8;桥0.2]

Table 1: Hidden states of boundaries.

Models	Boundary Enhance						
BERT	南	##京	##市	长	##江	##大	##桥
Labels aware	南B	京M	市E	长B	江M	大M	桥E
Word pos	南0	京1	市2	长0	江1	大2	桥3

Table 2: Input of boundary enhance.

TVect/ChinNER/tree/master/models/bert_ner) and Lattice LSTM(https://github.com/jiesutd/Lattice). Most implementation details followed those, including character and word embedding sizes, dropout, embedding initialization, and layer number. Additionally, the hidden size was set to 200 for small datasets Weibo and Resume, and 300 for larger datasets OntoNotes and MSRA. The initial learning rate was set to 0.005 for Weibo and 0.0015 for the rest three datasets.

4.2 Effectiveness Study

We conducted experiments on the four datasets to further verify the effectiveness of LSTM-HSB in combination with pre-trained model. Tables 3−6[1] show results on the MSRA, OntoNotes, Resume and Weibo datasets respectively against the compared baselines.

In Tables 3−6, compared methods include the best statistical models on these data set, which leveraged rich handcrafted features (Chen et al., 2006; Zhang et al., 2006; Zhou et al., 2013), character embedding features (Lu et al., 2016; Peng and Dredze, 2016a), radical features (Dong et al., 2016), cross-domain data, semi-supervised data (He and Sun, 2017b) and incorporating word lexicons methods (Zhang and Yang, 2018; Peng et al., 2019; Li et al., 2020). From the tables, we can see that the

[1]In Table 3−6, ∗ indicates that the model uses external labeled data for semi-supervised learning. † means that the model also uses discrete features.

Models	**P**	**R**	**F1**
Chen et al. (2006)	91.22	81.71	86.20
Zhang et al. (2006)*	92.20	90.18	91.18
Zhou et al. (2013)	91.86	88.75	90.28
Lu et al. (2016)	-	-	87.94
Dong et al. (2016)	91.28	90.62	90.95
Ma et al. (2020)* †	**94.63**	92.70	**93.66**
Li et al. (2020)* †	92.46	93.77	93.11
BiLSTM+CRF	92.23	90.52	91.37
BiLSTM-HSB+CRF	92.26	90.68	91.46
Lattice LSTM	93.68	92.33	93.00
Lattice LSTM-HSB	**93.82**	**93.26**	**93.54**
Bert+BiLSTM+CRF	94.72	94.10	94.41
Bert+BiLSTM-HSB+CRF	**95.08**	**94.51**	**94.79**

Table 3: Performance on MSRA.

performance of the LSTM-HSB method is better than baseline methods on four datasets. The average performance of the LSTM-HSB method is near to SOTA on four datasets. The reason of cannot over SOTA may be the embedding in static state and depending on tokenizations which may fail to recognize unnamed words like '江大桥(Daqiao Jiang)'. Comparing with BiLSTM+CRF and Lattice LSTM, we find that, the BiLSTM-HSB+CRF and Lattice LSTM-HSB methods have better performance. Those results show our method is transferable to different sequence-labeling architecture.

The proposed methods (Zhang and Yang, 2018; Ma et al., 2020) employ lattice-LSTM and consider the multiple tokenizations. The real difference be-

Models	P	R	F1
Yang et al. (2016)	65.59	71.84	68.57
Yang et al. (2016)*†	72.98	**80.15**	**76.40**
Che et al. (2013)*	**77.71**	72.51	75.02
Wang et al. (2013)*	76.43	72.32	74.32
Ma et al. (2020)*†	77.13	75.22	76.16
Li et al. (2020)*†	74.73	76.70	75.70
BiLSTM+CRF	74.60	74.91	74.75
BiLSTM-HSB+CRF	75.54	73.96	74.75
Lattice LSTM	75.51	**75.90**	75.70
Lattice LSTM-HSB	**76.42**	75.12	**75.77**
Bert+BiLSTM+CRF	77.27	79.41	78.33
Bert+BiLSTM-HSB+CRF	**77.72**	**79.74**	**78.71**

Table 4: Performance on OntoNotes.

Models	P	R	F1
Zhang and Yang (2018)*	93.72	93.44	93.58
Zhu and Wang (2019)	94.07	94.42	94.24
Liu et al. (2019)*	93.66	93.31	93.48
Ding et al. (2019)	94.53	94.29	94.41
Ma et al. (2020)*†	**96.14**	94.72	95.43
Li et al. (2020)*†	95.71	95.77	**95.74**
BiLSTM+CRF	94.49	94.49	94.49
BiLSTM-HSB+CRF	94.64	94.78	94.71
Lattice LSTM	**94.81**	94.11	94.46
Lattice LSTM-HSB	94.67	**94.72**	**94.69**
Bert+BiLSTM+CRF	94.71	95.98	95.34
Bert+BiLSTM-HSB+CRF	**95.42**	**96.27**	**95.85**

Table 5: Performance on Resume.

Models	NE	NM	Overall
Peng and Dredze (2015)	51.96	61.05	56.05
Peng and Dredze (2016a)*	55.28	62.97	58.99
He and Sun (2017a)	50.60	59.32	54.82
He and Sun (2017b)*	54.50	62.17	58.23
Ma et al. (2020)*†	58.12	64.20	59.81
Li et al. (2020)*†	**61.67**	65.27	**63.42**
BiLSTM+CRF	52.98	60.59	56.59
BiLSTM-HSB+CRF	53.08	61.48	57.85
Lattice LSTM	53.37	62.03	58.10
Lattice LSTM-HSB	**55.17**	**64.07**	**59.90**
Bert+BiLSTM+CRF	67.83	67.65	67.96
Bert+BiLSTM-HSB+CRF	**70.74**	**68.21**	**69.70**

Table 6: Performance on Weibo. NE, NM and Overall denote F1 scores for named entities, nominal entities (excluding named entities) and both, respectively.

tween this and the proposed methods should be discussed. However, they incorporate many wrong word lexicons without considering the whole sentences for segmentation. For example, in the sentence "南京市长江大桥(Nanjing Yangtze River Bridge)", they will incorporate wrong word '京市(Jing City)' without considering the whole sentence for segmentation. In our method, we consider the whole sentence for hidden boundary states. The smoothing technique to entity boundaries is better than hard word incorporation to show sentence structure and relationship between nearby characters.

For Chinese NER, the hidden boundary states in LSTM are unbalanced of each character. We activate hidden states of character with context for boundary. It can be trained fast and reduce the parameters in model. During the hidden boundary of sentence, we provide a soft way to locate the word boundary. It simplifies the model to learn structure from large data. With the additional states of boundary, the model can fast and better learn the structure of sentence.

4.3 Compatibility with BERT

We compare LSTM-HSB with BERT on four datasets. We download the specified pretrained BERT model provided by huggingface. We use bert-base-chinese (`https://storage.googleapis.com/bert_models/2018_11_03/chinese_L-12_H-768_A-12.zip`) for Chinese task. In these experiments, we use BERT encoders to obtain the character representations.

From the Table 3−6, we can see that the LSTM-HSB method with BERT outperforms the BERT tagger on all four datasets. These results show that the LSTM-HSB method can be effectively combined with pre-trained model. Moreover, the results also verify the effectiveness of our method in utilizing lexicon information, which means it can complement the information obtained from the pre-trained model. We also find that, Bert+BiLSTM-HSB+CRF have an improvement over Bert+BiLSTM+CRF. Those results show our method is transferable to sequence-labeling architecture and improve the F1 in Chinese NER with pre-trained model.

4.4 Ablation Study

To investigate the contribution of each component of our method, we conduct ablation experiments on all four datasets, as shown in table 7.

In the "LSTM-HSB w/ tanh" experiment, we replace the sigmoid activation as tanh activation in LSTM-HSB.

$$H_t = h_t + tanh(W^h_{tanh} h_t + b^h_{tanh}), \quad (4)$$

Models	MSRA	OntoNotes	Resume	Weibo
LSTM-HSB	91.46	74.75	94.71	57.85
LSTM-HSB w/ tanh	88.69	72.06	91.75	53.44
LSTM-HSB w/ cell-sigmoid	90.76	73.19	92.91	54.80
LSTM-HSB w/ cell-tanh	87.61	69.95	90.72	52.17

Table 7: An ablation study of the proposed model.

where W^h_{tanh} and b^h_{tanh} are trainable parameters.

The degradation in performance on all four datasets indicates the importance of the sigmoid activation, and confirms the advantage of our method. The information of boundary should be activate as values of gate rather than tanh input. The values of [0,1] is better than [-1,1] to represent boundaries between nearby characters.

In the "LSTM-HSB w/ cell-sigmoid" experiment, we replace the hidden states as cell states in LSTM-HSB.

$$H_t = h_t + \sigma(W^C C_t + b^C), \quad (5)$$

where W^C and b^C are trainable parameters.

The degradation in performance on all four datasets indicates the importance of the hidden states, and confirms the advantage of our method. The cell states can be activated as boundaries but perform not better than hidden states which encode union structure with parameters in LSTM.

In the "LSTM-HSB w/ cell-tanh" experiment, we replace the hidden states as cell states and sigmoid activation as tanh activation in LSTM-HSB.

$$H_t = h_t + tanh(W^C_{tanh} C_t + b^C_{tanh}). \quad (6)$$

where W^C_{tanh} and b^C_{tanh} are trainable parameters.

The degradation in performance on all four datasets indicates the importance of the hidden states and the sigmoid activation, and confirms the advantage of our method. The combines of boundary gates and hidden states are harmonious in model.

5 Conclusion

In this work, we address the hidden boundary state in Chinese NER. We propose a novel method to locate the soft boundary with considering the sequence of characters in whole sentence, which reduces many wrong words incorporated into the character representations. We use LSTM with hidden state activation instead of CWS system to embed the word-lever information. Experimental studies show that our performances have an improvement of existing methods.

References

Shuyang Cao and Lu Wang. 2022. HIBRIDS: Attention with hierarchical biases for structure-aware long document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 786–807, Dublin, Ireland. Association for Computational Linguistics.

Wanxiang Che, Mengqiu Wang, Christopher D Manning, and Ting Liu. 2013. Named entity recognition with bilingual constraints. In *NAACL*, pages 52–62.

Aitao Chen, Fuchun Peng, Roy Shan, and Gordon Sun. 2006. Chinese named entity recognition with conditional probabilistic models. In *SIGHAN Workshop on Chinese Language Processing*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ruixue Ding, Pengjun Xie, Xiaoyan Zhang, Wei Lu, Linlin Li, and Luo Si. 2019. A neural multi-digraph model for Chinese NER with gazetteers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1462–1467, Florence, Italy. Association for Computational Linguistics.

Chuanhai Dong, Jiajun Zhang, Chengqing Zong, Masanori Hattori, and Hui Di. 2016. Character-based lstm-crf with radical-level features for chinese named entity recognition. In *Natural Language Understanding and Intelligent Applications*, pages 239–250. Springer.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. 2021. SpanNER: Named entity re-/recognition as span prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7183–7195, Online. Association for Computational Linguistics.

Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuan-Jing Huang. 2019. A lexicon-based graph neural network for chinese ner. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1039–1049.

Hangfeng He and Xu Sun. 2017a. F-score driven max margin neural network for named entity recognition in chinese social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 713–718.

Hangfeng He and Xu Sun. 2017b. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Jingzhou He and Houfeng Wang. 2008. Chinese named entity recognition and word segmentation based on character. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

Wu Hong, Zhuosheng Zhang, Jinyuan Wang, and Hai Zhao. 2022. Sentence-aware contrastive learning for open-domain passage retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1062–1074, Dublin, Ireland. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *ArXiv*, abs/1508.01991.

Gina-Anne Levow. 2006. The third international chinese language processing bakeoff: Word segmentation and named entity recognition. In *SIGHAN Workshop on Chinese Language Processing*, pages 108–117.

Fei Li, Zhichao Lin, Meishan Zhang, and Donghong Ji. 2021a. A span-based model for joint overlapped and discontinuous named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4814–4828, Online. Association for Computational Linguistics.

Fei Li, Zheng Wang, Siu Cheung Hui, Lejian Liao, Dandan Song, Jing Xu, Guoxiu He, and Meihuizi Jia. 2021b. Modularized interaction network for named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 200–209, Online. Association for Computational Linguistics.

Haibo Li, Masato Hagiwara, Qi Li, and Heng Ji. 2014. Comparison of the impact of word segmentation on name tagging for chinese and japanese. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jing Li, Aixin Sun, and Yukun Ma. 2021c. Neural named entity boundary detection. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1790–1795.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: Chinese NER using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6836–6842, Online. Association for Computational Linguistics.

Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Fangzheng Xu, Huan Gui, Jian Peng, and Jiawei Han. 2018. Empower sequence labeling with task-aware neural language model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Wei Liu, Tongge Xu, Qinghua Xu, Jiayu Song, and Yueran Zu. 2019. An encoding strategy based word-character LSTM for Chinese NER. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2379–2389, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhangxun Liu, Conghui Zhu, and Tiejun Zhao. 2010. Chinese named entity recognition with a sequence labeling approach: based on characters, or based on words? In *Advanced intelligent computing theories and applications. With aspects of artificial intelligence*, pages 634–640. Springer.

Yanan Lu, Yue Zhang, and Dong-Hong Ji. 2016. Multi-prototype chinese character embedding. In *LREC*.

Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, and Xuanjing Huang. 2020. Simplify the usage of lexicon in Chinese NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5951–5960, Online. Association for Computational Linguistics.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2022. Structural characterization for dialogue disentanglement. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 285–297, Dublin, Ireland. Association for Computational Linguistics.

Xinnian Mao, Yuan Dong, Saike He, Sencheng Bao, and Haila Wang. 2008. Chinese word segmentation and named entity recognition based on conditional random fields. In *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*.

Minlong Peng, Ruotian Ma, Qi Zhang, and Xuanjing Huang. 2019. Simplify the usage of lexicon in chinese ner. *ArXiv*, abs/1908.05969.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554.

Nanyun Peng and Mark Dredze. 2016a. Improving named entity recognition for chinese social media with word segmentation representation learning. In *ACL*, page 149.

Nanyun Peng and Mark Dredze. 2016b. Learning word segmentation representations to improve named entity recognition for chinese social media. *CoRR*, abs/1603.00786.

Richard Sproat, Chilin Shih, William Gale, and Nancy Chang. 1994. A stochastic finite-state word-segmentation algorithm for Chinese. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 66–73, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Jinsong Su, Jiali Zeng, Deyi Xiong, Yang Liu, Mingxuan Wang, and Jun Xie. 2018. A hierarchy-to-sequence attentional neural machine translation model. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3):623–632.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2019. Leverage lexical knowledge for chinese named entity recognition via collaborative graph network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3821–3831.

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

Mengqiu Wang, Wanxiang Che, and Christopher D Manning. 2013. Effective bilingual constraints for semi-supervised learning of named entity recognizers. In *AAAI*.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Lu Xu, Zhanming Jie, Wei Lu, and Lidong Bing. 2021. Better feature integration for named entity recognition. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3457–3469, Online. Association for Computational Linguistics.

Jie Yang, Zhiyang Teng, Meishan Zhang, and Yue Zhang. 2016. Combining discrete and neural features for sequence labeling. In *CICLing*. Springer.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Suxiang Zhang, Ying Qin, Juan Wen, and Xiaojie Wang. 2006. Word segmentation and named entity recognition for sighan bakeoff3. In *SIGHAN Workshop on Chinese Language Processing*, pages 158–161.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), 1554-1564.*

Junsheng Zhou, Weiguang Qu, and Fen Zhang. 2013. Chinese named entity recognition via joint identification and categorization. *Chinese journal of electronics*, 22(2):225–230.

Enwei Zhu and Jinpeng Li. 2022. Boundary smoothing for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7096–7108, Dublin, Ireland. Association for Computational Linguistics.

Yuying Zhu and Guoxin Wang. 2019. CAN-NER: Convolutional Attention Network for Chinese Named Entity Recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3384–3393, Minneapolis, Minnesota. Association for Computational Linguistics.