Computer Science and Application 计算机科学与应用, 2020, 10(7), 1359-1366
Published Online July 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.107140

Hans汉斯

Ancient Chinese Named Entity Recognition Based on Deeping Learning

Macuo Zhuo[1,2], Duanzhu Sangjie[1,2], Rangjia Cai[1,2]

[1]Key Laboratory of Tibetan Information Processing of Ministry of Education, Qinghai Normal University, Xining Qinghai
[2]Qinghai Provincial Key Laboratory of Tibetan Information Processing and Machine Translation, Xining Qinghai
Email: 2353498508@qq.com

Received: Jul. 2nd, 2020; accepted: Jul. 16th, 2020; published: Jul. 24th, 2020

Abstract

Named entity recognition is one of the basic tasks of natural language processing. At present, the research on Chinese named entities recognition is mostly for modern Chinese, and the research on it for ancient Chinese is less involved. So in this paper, taking the *War State Policy* as an example and according to the characteristics of ancient Chinese text, we use the Lattice Long and Short-Term memory (Lattice LSTM) neural network to construct a named entity recognition model to solve the problem of information extraction of ancient Chinese. Experiment result shows that Lattice LSTM can learn to automatically find all the dictionary-matched words from the context to achieve better named entity recognition performance. The F1 value reaches 92.16%.

Keywords

Neural Network Model, Ancient Chinese, Named Entity Recognition, Conditional Random Field

基于深度学习的古汉语命名实体识别研究

卓玛措[1,2]，桑杰端珠[1,2]，才让加[1,2]

[1]青海师范大学藏文信息处理教育部重点实验室，青海 西宁
[2]青海省藏文信息处理与机器翻译重点实验室，青海 西宁
Email: 2353498508@qq.com

收稿日期：2020年7月2日；录用日期：2020年7月16日；发布日期：2020年7月24日

摘 要

命名实体识别是自然语言处理的基础任务之一。而目前中文命名实体识别研究大多是面向现代汉语的，针对古汉语的这方面研究工作涉及较少。因此，本文以《战国策》为例，根据古汉语独特的子语言特征，

文章引用: 卓玛措, 桑杰端珠, 才让加. 基于深度学习的古汉语命名实体识别研究[J]. 计算机科学与应用, 2020, 10(7): 1359-1366. DOI: 10.12677/csa.2020.107140

利用网格长短期记忆(Lattice LSTM)神经网络构建命名实体识别模型以解决古汉语中的信息提取问题。实验结果表明，Lattice LSTM能够学会从语境中自动找到所有与词典匹配的词汇，以取得较好的命名实体识别性能。实验结果中的F1值达到92.16%。

关键词

神经网络模型，古汉语，命名实体识别，条件随机场

1. 引言

命名实体识别[1] (Named Entity Recognition, NER)是自然语言处理(Natural Language Processing, NLP)基础性工作之一，它可以准确地从文本中识别出人名、机构名、地名、时间等信息，为信息检索、机器翻译、舆情分析等下游自然语言处理任务提供重要的特征信息。过去，命名实体识别任务多采用基于规则的方法、基于统计的方法和基于规则和统计相结合的方法[2]。

近年来，深度神经网络在自然语言处理领域广泛地受到关注，与上述方法相比，基于深度神经网络的方法具有更强的泛化能力、对人工特征依赖较少的优点。因此，面向现代汉语和英语等大语种，研究者已提出了许多基于深度神经网络的命名实体识别模型[3]-[18]，但针对古汉语在这方面的研究才刚刚起步。

鉴于此，本文以《战国策》为例，根据古汉语独特的子语言特征，利用网格长短期记忆(Lattice LSTM)神经网络构建命名实体识别模型以解决古汉语中的信息提取问题。该方法将传统的 LSTM 单元改进为网格 LSTM，在字模型的基础之上显性利用词与词序信息，从而避免了分词错误传递的问题。实验结果表明，Lattice LSTM 能够学会从语境中自动找到所有与词典匹配的词汇，以取得较好的命名实体识别性能。在本研究构建的数据集上 F1 值达到 92.16%。

2. 模型

在英文领域，第一个采用神经网络进行命名实体识别的是 Hammerton 等人，由于 LSTM 良好的序列建模能力，LSTM-CRF [19]模型成为命名实体识别的基础架构之一，很多方法都是以 LSTM-CRF 为主体框架，在此之上融入各种相关特征。本文将 LSTM-CRF 作为主要网络结构，并且在该模型对一系列输入字符进行编码的同时将所有与词典匹配的词汇网格结构融入模型中。

一般将输入序列表示为 $s=c_1,c_2,\cdots,c_n$。其中，c_j 代表第 j 个字符。本文中应用 $t(i,k)$ 表示索引 j，代表第 i 个词的第 k 个字符。比如“医扁鹊”，索引从 1 开始，那么 $t(1,1)=1$(医)，$t(2,1)=2$(扁)。本研究运用 BIO 标注策略进行字粒度和词粒度的命名实体识别标注，古汉语命名实体识别的字序列和标记序列举例说明如表 1 所示。

Table 1. Examples of character sequence and label sequence in ancient Chinese named entity recognition
表 1. 古汉语命名实体识别的字序列和标记序列举例

字	舜	虽	贤	不	遇	尧	不	得	为	天	子
标记	B-PER	O	O	O	O	B-PER	O	O	O	B-POS	I-POS

注：“O”是实体外部标记，“B”是实体开始标记，“I”是实体内部标记，“PER”表示人名，“POS”表示官职。

2.1. 基于字的模型

基于字的命名实体识别(见图 1)存在一种缺陷，即无法充分利用词的显性以及词序信息。

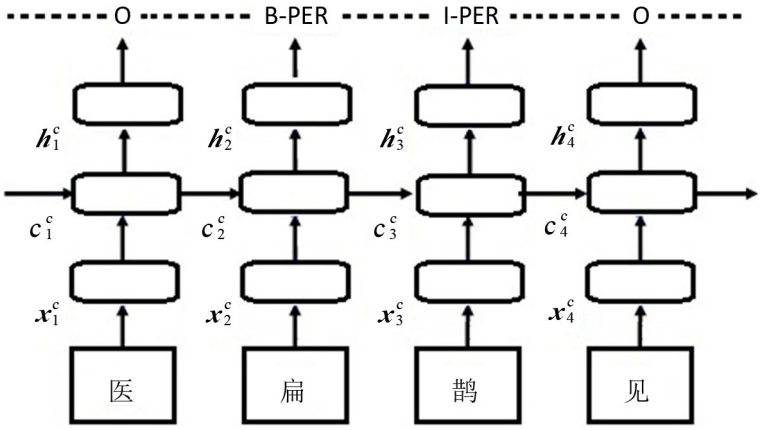

Figure 1. Character-based model
图 1. 基于字符的模型

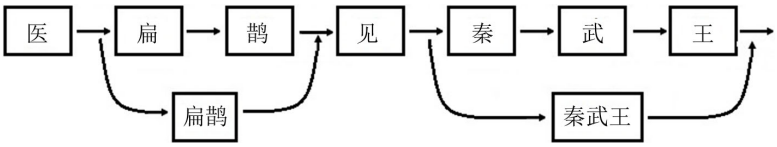

Figure 2. Word-character lattice
图 2. 词－字符网格

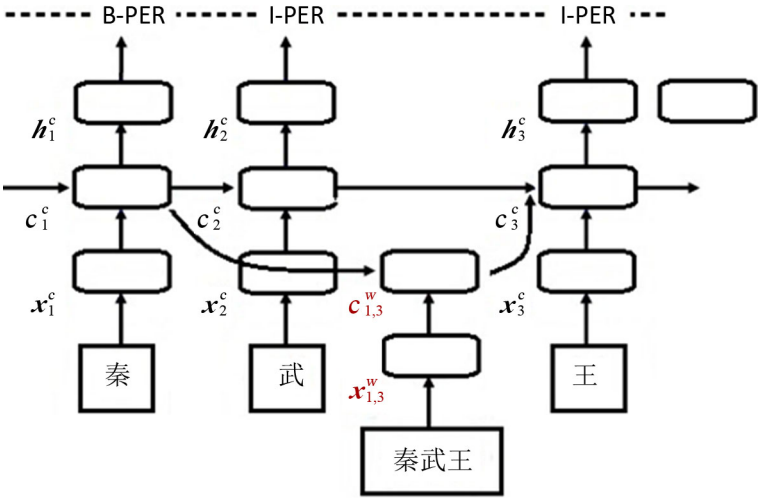

Figure 3. Lattice model
图 3. Lattice 模型

2.2. 基于词的模型

基于词的模型见图 4。

2.3. Lattice LSTM 模型

本文利用 Lattice LSTM [20]来处理句子中的词汇词(lexicon word)，从而将所有潜在词信息全部整合

到基于字符的LSTM-CRF中，见图2。并使用一个自动获取的词典来匹配句子，进而构建基于词的Lattice，见图3。由于在网格中存在指数数量的单词到字符路径，因此使用Lattice LSTM结构来自动控制从句子的开头到结尾的信息流。门控单元用于将不同路径的信息动态的传输到每个字符。在训练数据集上训练后，Lattice LSTM能够学会从信息流中自动找到有用的词，从而提升命名实体识别性能，见图5。与基于字符和基于词的命名实体识别方法相比，本文采用的模型优势在于利用词汇的显性信息进行分词，而不是仅仅自动关注，从而减少分词误差。

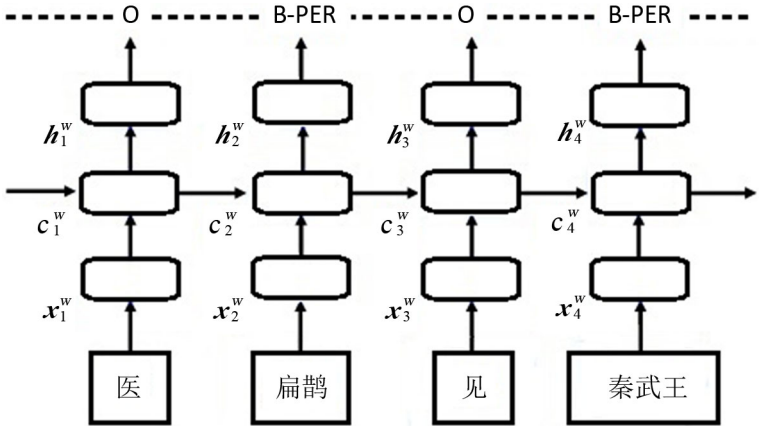

Figure 4. Word-based model
图4. 基于词的模型

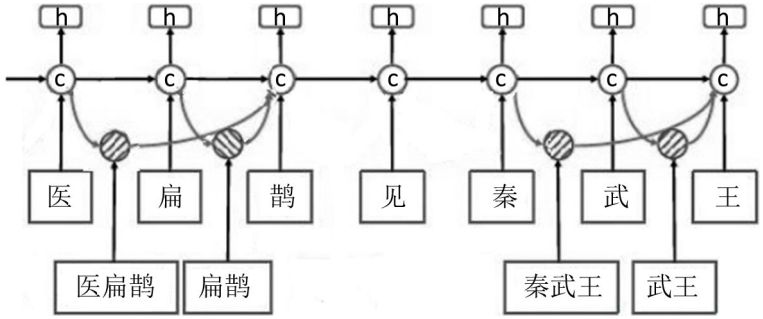

Figure 5. Lattice LSTM model
图5. Lattice LSTM模型

2.4. LSTM层

RNN循环神经网络理论上可以处理任意长度的序列信息，但实际应用中，当序列过长时会出现梯度消失的问题，且很难学到长期依赖的特征。因此，Graves 等人[21]改进了循环神经网络，提出长短期记忆网络(Long Short-Term Memory) LSTM模型。LSTM单元通过输入门、遗忘门和输出门来控制信息传递。它是一种特殊的RNN，能够学习长期的规律，应用十分广泛。LSTM编码单元如图6所示。

具体计算过程如公式(1)~(6)所示：

$$i_t = \sigma\left(W_i h_{t-1} + U_i + b_i\right) \tag{1}$$

$$f_t = \sigma\left(W_f h_{t-1} + U_f x_t + b_f\right) \tag{2}$$

$$\widetilde{c}_t = \tanh\left(W_c h_{t-1} + U_c x_t + b_f\right) \tag{3}$$

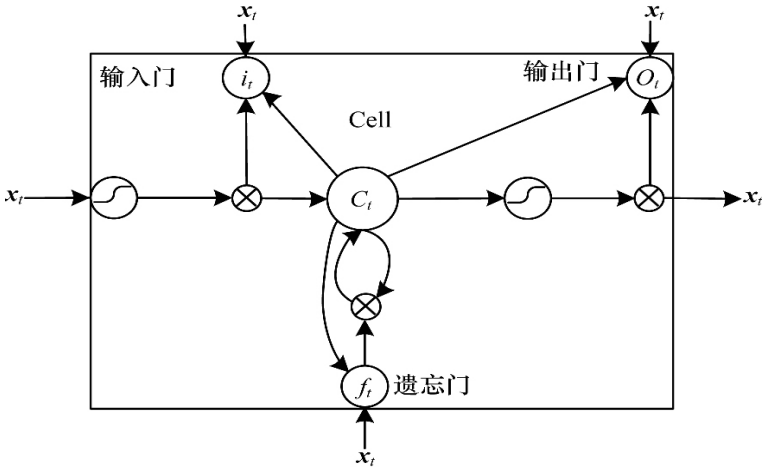

Figure 6. LSTM unit
图 6. LSTM 编码单元

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{4}$$

$$o_t = \sigma\left(W_o h_{t-1} + U_o x_t + b_o\right) \tag{5}$$

$$h_t = o_t \odot \tanh\left(c_t\right) \tag{6}$$

其中，σ 是 sigmoid 函数，$\odot$ 是点积。x_t 为时刻 t 的输入向量，h_t 是隐藏状态，也是输出向量，包含前面 t 时刻所有有效信息。c_t 是一个更新门，控制信息流入下一个时刻；f_t 是一个遗忘门，控制信息丢失；二者共同决定隐藏状态的输出。

3. 实验及结果分析

3.1. 实验数据

由于目前古汉语命名实体识别缺乏公开的标注数据集，因此本文人工构建了一个古汉语命名实体识别数据集。该数据集包括训练集、开发集、测试集，训练集共包含 43.995 K 个字，开发集包含 5.843 K 个字，测试集包含 5.849 K 个字。各类实体统计如表 2 所示。

Table 2. Statistics of entity number
表 2. 实体个数统计

数据集	语料数量	人名数量	地名数量	职官数量
训练集	43,995	1,255	2,671	101
开发集	5,843	125	285	37
测试集	5,849	144	315	26

3.2. 标注策略与评价指标

命名实体识别的标注策略有 BIO 模式，BIOE 模式，BIOES 模式。本文采用的是 BIO 标注策略，其中 B 表示实体开始，I 表示实体非开始部分。O 表示不是实体的部分。在预测实体边界的时候需要同时预测实体类型，所以待预测的标签一共 7 种，分别是 O，B-PER，I-PER，B-LOC，I-LOC，B-POS，I-POS。

在测试过程中，只有当一个实体的边界和实体的类型完全正确时，才判断该实体预测正确。

命名实体识别的评价指标有精确率(P)、召回率(R)和 $F1$ 值。具体定义如公式(7)：T_p 为模型识别正确的实体个数，F_p 为模型识别到的不相关实体个数，F_n 为相关实体但是模型没有检测到的个数。

$$P = \frac{T_P}{T_P + F_P} \times 100\%$$

$$R = \frac{T_P}{T_P + F_n} \times 100\% \tag{7}$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\%$$

3.3. 实验环境与超参设置

本研究中的实验环境为 Python3.6，深度学习框架为 Pytorch1.4.0 神经网络超参的取值会影响神经网络的性能。本文的神经网络参数设定如表 3 所示。

Table 3. Neural network hyperparameter values
表 3. 神经网络超参取值

参数	取值	参数	取值
音节向量维度	50	词向量维度	50
Lattice 向量维度	50	Lattice 丢弃率	0.5
丢弃率(Dropout)	0.5	学习率(lr)	0.05
LSTM 层	1	主 LSTM 隐藏层维度	200

3.4. 实验设计与结果

为了验证本研究中所使用的模型对古汉语命名实体识别数据集中的人名、地名、官职三大类实体的识别性能，本文分别采用三种神经网络模型设计了三个实验。其中主要实验模型为 Lattice LSTM，对比实验模型为 BiLSTM-CRF 和 BiLSTM-CNN-CRF。实验的评价指标有准确率(P)、召回率(R)和综合指标 $F1$ 值。各模型实验结果见表 4。

Table 4. Experiment results of each model (%)
表 4. 各模型的实验结果(%)

模型	准确率	召回率	$F1$ 值
BiLSTM-CRF	88.30	87.70	87.92
BiLSTM-CNN-CRF	89.50	89.10	89.25
Lattice LSTM	**92.42**	**91.90**	**92.16**

实验结果表明 Lattice LSTM 模型能有效提升实体识别的性能。各模型随着训练轮数 F1 值变化如图 7 所示。

3.5 实体识别实例

以本研究构建的数据集中的一个句子为例展示 Lattice LSTM 模型的实体识别效果。具体实例如表 5

所示。

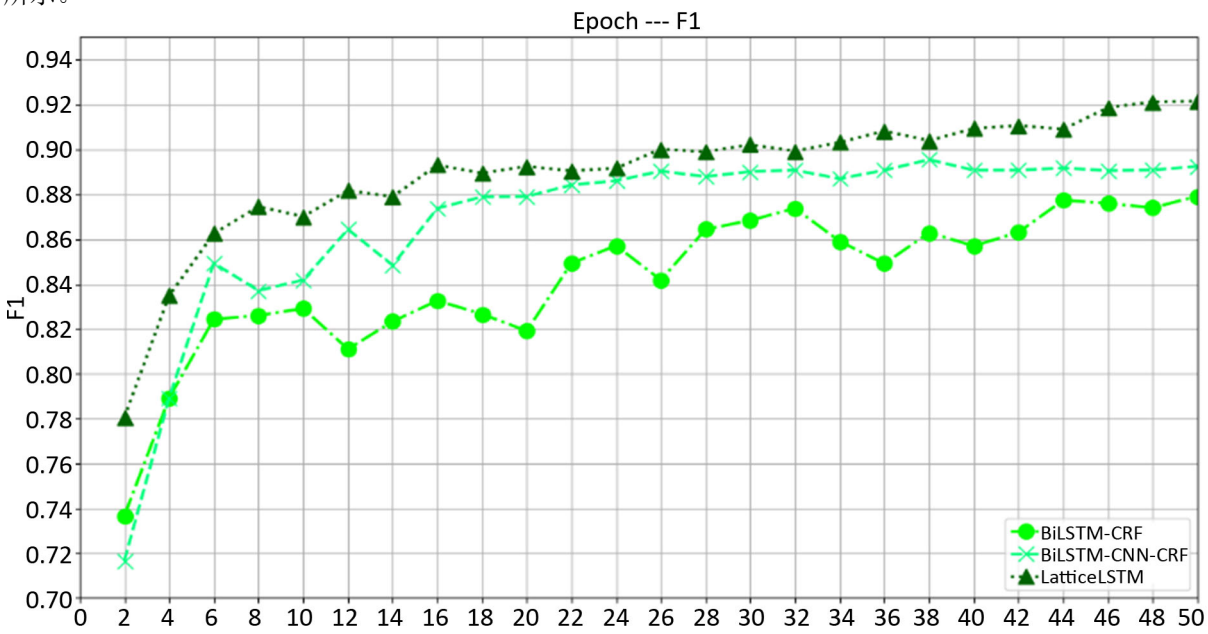

Figure 7. F1 values
图 7. F1 值变化图

Table 5. Examples of ancient Chinese entity recognition
表 5. 古汉语实体识别实例

句子	故以舜汤武不遭时不得帝王
正确的分词	故以舜汤武不遭时不得帝王
自动分词	故以舜汤武不遭时不得帝王
Lattice 分词	故以舜汤汤武舜汤武不遭遭时不遭时不得帝王
BiLSTM-CRF	故以舜汤武 $_{PER}$ 不遭时不得**帝王** $_{POS}$
BiLSTM-CNN-CRF	故以舜汤武 $_{PER}$ 不遭时不得**帝王** $_{POS}$
Lattice LSTM	故以**舜** $_{PER}$ **汤** $_{PER}$ **武** $_{PER}$，不遭时不得**帝王** $_{POS}$

注：斜体表示识别不正确的实体，粗体表示识别正确的实体。

4. 结束语

针对古汉语命名实体识别所面临的问题，本文采用了一种同时关注字信息和词信息进行实体识别的深度学习模型。该模型将传统的 LSTM 单元改进为网格 LSTM，在字符模型的基础之上显性利用词和词序信息，从而避免了分词错误传递的问题；利用具有长短期记忆功能的 LSTM 模型作为隐藏层，可以解决古汉语文本中部分实体结构较长的问题；最后使用 CRF 作为标签推理层以解决文本序列标签依赖问题。在已构建的古汉语命名实体识别数据集上进行实验，实验结果证明了 Lattice LSTM 模型的有效性。

今后，本文的研究工作应该在数据和词典规模方面加大力度，从而进一步提高模型的整体性能。另外，还应该针对古汉语文本进行广泛深入的语言信息处理方面的研究，以便获得更多有价值的知识。

基金项目

国家自然科学基金(61662061, 61063033)、国家重点研发计划(2017YFB1402200)、青海省科技厅项目

(2015-SF-520)。

参考文献

[1] Hammerton, J. (2003) Named Entity Recognition with Long Short-Term Memory. In: *Conference on Natural Language Learning*, Association for Computational Linguistics, Stroudsburg, 172-175. https://doi.org/10.3115/1119176.1119202

[2] 张海楠, 伍大勇, 刘悦, 等. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报, 2017, 31(4): 28-35.

[3] Ma, X. and Hovy, E. (2016) End-to-End Sequence Labeling via Bi-Directional LSTM-CNNs-CRF. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics*, Volume 1: Long Papers, Berlin, August 2016, 1064-1074. https://doi.org/10.18653/v1/P16-1101

[4] Chiu, J.P.C. and Nichols, E. (2016) Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, **4**, 357-370. https://doi.org/10.1162/tacl_a_00104

[5] Lample, G., Ballesteros, M., Subramanian, S., *et al*. (2016) Neural Architectures for Named Entity Recognition. *Proceedings of NAACL-HLT* 2016, San Diego, 12-17 June 2016, 260-270. https://doi.org/10.18653/v1/N16-1030

[6] Nothman, J., Ringland, N., Radford, W., Murphy, T. and Curran, J.R. (2013) Learning Multilingual Named Entity Recognition from Wikipedia. *Artificial Intelligence*, **194**, 151-175.

[7] Santos, C.N.D. and Guimarães, V. (2015) Boosting Named Entity Recognition with Neural Character Embeddings.

[8] 买买提阿依甫, 吾守尔·斯拉木, 帕丽旦·木合塔尔, 等. 基于 BiLSTM-CNN-CRF 模型的维吾尔文命名实体识别[J]. 计算机工程, 2018, 44(8): 230-236.

[9] 王洁, 张瑞东, 吴晨生. 基于 GRU 的命名实体识别方法[J]. 计算机系统应用, 2018, 27(9): 18-24.

[10] 李丽双, 郭元凯. 基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别[J]. 中文信息学报, 2018, 32(1): 116-122.

[11] 周晓磊, 赵薛蛟, 刘堂亮, 宗子潇, 王其乐, 里剑桥. 基于 SVM-BiLSTM-CRF 模型的财产纠纷命名实体识别方法[J]. 计算机系统应用, 2019, 28(1): 245-250.

[12] 杨文明, 褚伟杰. 在线医疗问答文本的命名实体识别[J]. 计算机系统应用, 2019, 28(2): 8-14.

[13] He, J. and Wang, H. (2008) Chinese Named Entity Recognition and Word Segmentation Based on Character. *Proceedings of the Sixth SIGHAN Workshop on Chinese Language Processing*, Hyderabad, 11-12 January 2008, 128-132.

[14] Liu, Z., Zhu, C. and Zhao, T. (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*, Springer, Berlin Heidelberg, 634-640. https://doi.org/10.1007/978-3-642-14932-0_78

[15] 杨培, 杨志豪, 罗凌, 林鸿飞, 王健. 基于注意机制的化学药物命名实体识别[J]. 计算机研究与发展, 2018, 55(7): 1548-1556.

[16] Li, H., Hagiwara, M., Li, Q., *et al*. (2014) Comparison of the Impact of Word Segmentation on Name Tagging for Chinese and Japanese. 2014 *LREC*, Reykjavik, 26-31 May 2014, 2532-2536.

[17] Chen, W., Zhang, Y. and Isahara, H. (2006) Chinese Named Entity Recognition with Conditional Random Fields. *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, Sydney, 22-23 July 2006, 118-121.

[18] Dong, C., Zhang, J., Zong, C., *et al*. (2016) Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition. In: *Natural Language Understanding and Intelligent Applications*, Springer, Cham, 239-250. https://doi.org/10.1007/978-3-319-50496-4_20

[19] Huang, Z., Xu, W. and Yu, K. (2015) Bidirectional LSTM-CRF Models for Sequence Tagging.

[20] Zhang, Y. and Yang, J. (2018) Chinese NER Using Lattice LSTM. *Proceedings of the* 56*th Annual Meeting of the Association for Computational Linguistics*, Volume 1: Long Papers, Melbourne, July 2018, 1554-1564. https://doi.org/10.18653/v1/P18-1144

[21] Graves, A. (2013) Generating Sequences with Recurrent Neural Networks. *Computer Science*, 1-43.