Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2021, 10(3), 215-223
Published Online August 2021 in Hans. http://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2021.103021

Hans汉斯

基于BERT的中文命名实体识别方法

王　希，张传武，刘东升

西南民族大学电子信息学院，四川 成都
Email: 1322108854@qq.com

收稿日期：2021年6月18日；录用日期：2021年7月2日；发布日期：2021年7月22日

摘　要

由于中文与英文本身存在较大的差异，中文命名实体识别的研究存在一系列的挑战。目前来说，BLSTM-CRF模型使用最为广泛。该模型采用深度学习模型与统计模型相结合的方式进行中文命名实体识别，能够有效提取出文本中的上下文信息并考虑标签之间的关系。但由于中文存在多义字或词，存在一个句子中相同字词含义差别很大的情况，该模型在这种情况下实体识别的性能并不理想。为了更好地实现字表示既可以包含各种多样化的句法和语义表示，又可以对多义字进行建模，引入了BERT语言模型，此模型可以根据上下文信息计算出更高的全局性字词向量表示以及在句中的权重。BERT-BLSTM-CRF命名实体识别模型通过BERT预训练模型增强词向量的表示，BLSTM获取上下文语义标签序列，再使用CRF求得最优解。本文使用人民日报数据集对提出模型的进行实验测试，从实验结果可以发现，该模型的实体识别性能与传统模型相比有较大的提升。

关键词

命名实体识别，BERT，BLSTM，条件随机场

Chinese Named Entity Recognition Method Based on BERT

Xi Wang, Chuanwu Zhang, Dongsheng Liu

College of Electronics and Information, Southwest Minzu University, Chengdu Sichuan
Email: 1322108854@qq.com

Received: Jun. 18th, 2021; accepted: Jul. 2nd, 2021; published: Jul. 22nd, 2021

Abstract

Due to the large differences between Chinese and English, there are a series of challenges in the research of Chinese named entity recognition. Currently, the BLSTM-CRF model is the most widely

文章引用: 王希, 张传武, 刘东升. 基于 BERT 的中文命名实体识别方法[J]. 人工智能与机器人研究, 2021, 10(3): 215-223. DOI: 10.12677/airr.2021.103021

used. The model uses a combination of deep learning models and statistical models for Chinese named entity recognition, which can effectively extract contextual information in the text and consider the relationship between tags. However, due to the presence of polysemous characters or words in Chinese, the meaning of the same words in a sentence may be very different. In this case, the performance of the entity recognition of the model is not ideal. In order to better realize that the word representation can not only contain a variety of diversified syntax and semantic representations, but also can model polysemous words, the BERT language model is introduced, which can calculate higher global word vector representations and weights in sentences based on context information. The BERT-BLSTM-CRF named entity recognition model enhances the representation of word vectors through the BERT pre-training model, uses BLSTM to obtain contextual semantic label sequences, and then uses CRF to find the optimal solution. Using the People's Daily data set to test the proposed model, it can be found that the entity recognition performance of the model is greatly improved compared with the traditional model.

Keywords

Named Entity Recognition, BERT, BLSTM, Conditional Random Fields

1. 引言

随着人工智能技术的飞速发展和互联网数据的爆炸式增长，如何从海量的数据中迅速准确地提取关键的信息并挖掘其潜在的价值，是一项急需解决的任务。这通常需要用到自然语言处理技术(Natural Language Processing, NLP)，NLP 技术在人工智能产业中具有非常重要的地位，推动了人工智能持续发展和突破，而命名实体识别是 NLP 的一项基础却又重要的组成部分。命名实体识别的主要任务是从海量非结构文本信息中识别出所需要的实体信息，已被广泛应用于机器翻译、智能搜索、智能问答等领域。

近年来，基于神经网络的命名实体识别收到了广泛关注，国外对命名实体识别技术的研究已经取得了较好的成绩。Lample 等人[1]提出了一种新的基于神经网络的模型，获得了最先进的性能；Yang 等人[2]提出了一种可以在多任务和跨语言联合训练条件下提高性能的用于序列标注的深度层次递归神经网络；de Oliveira 等人[3]提出一种 FS-NER 方法，使用过滤器处理未标记的数据，可以以灵活的方式快速处理信息，比 CRF 监督学习方法更加实用。在国内，罗熹等人[4]运用融合领域词典的字符集特征表示方法，并应用于传统 BLSTM-CRF 模型，在中文临床命名实体识别任务中取得较优的识别性能；王昊等人[5]使用法律判决书文本作为数据集，通过与其他模型比较，得出 ALBERT-BLSTM-CRFs 效果更好且迁移能力更强的结论；钟诗胜等人[6]提出一种引入词集级注意力机制的中文命名识别方法，可以忽略不可靠的信息，从而提高识别性能。

目前中文命名实体识别存在很大的挑战，主要由于中文与英文本身存在较大的差异。原因有以下几点：其一，中文有字词之分，在句子中词之间没有用空格来划分开，这使得很多适合英文的命名实体识别方法无法在中文中使用；其二，自然语言处理的中文数据集相较英文数据集而言偏少；其三，中文存在一词多义，在不同的语境下，同一个词的意义可能有很大区别。所以，对中文命名实体识别的研究具有重大的意义。

2. 相关工作

命名实体识别最早是使用基于规则和字典的方法，这类方法根据人工事先定义的规则或词典对实体进行匹配抽取，准确率高，但面对大量数据集或全新领域时，需要重新建立新的规则库或词典，泛化性不高。

然后是基于统计机器学习的方法，这类方法依然需要利用人工标注语料作为训练集来训练机器学习模型，再利用训练好的模型实现预测[7]。这种方法的优点是标注的数据越多准确率越高，缺点是标注的成本过高[8]。

深度神经网络方法减少了模型对人工标注数据的依赖，可以构建不同特性的信息之间的关系，得到目标的特定表达，有大量的研究结果表明，将深度学习运用于命名实体识别任务中能够取得更好的结果[9]。Collobert 等人[10]提出一个多层神经网络架构，通过大量大部分未标记的训练数据来学习内部表示，实验证明该系统具有良好的性能。随后，出现了使用循环神经网络(Recurrent Neural Network, RNN)代替神经网络(Neural Network, NN)的模型，通常使用长短时记忆网络 LSTM (Long-Short Term Memory)。例如，Huang 等人[11]提出了 BLSTM (Bidirectional Long-Short Term Memory)与 CRF 结合的模型用于 NLP 序列标记；Chiu 等人[12]提出了包含一个双向 LSTM 和一个字符级的 CNN 的新神经网络结构，在少数特征工程情况下，在命名实体识别方面取得了最先进的结果；Ma 等人[13]结合双向 LSTM、CNN 和 CRF，提出了一种端到端的系统，无需特征工程，在大数据量的序列标记任务中表现非常好；Rei 等人[14]提出基于注意机制的字符级模型，在所有评价上都由于连接词级和字符级表征。

3. 命名实体识别模型

3.1. BLSTM 模型

在命名实体识别方法中，神经网络模型相较传统机器学习模型取得了更好的成绩。LSTM 在循环神经网络(Recurrent Neural Network, RNN)的基础上，将隐藏层的更新被专门构建的记忆单元所取代，可以解决 RNN 可能造成的梯度爆炸或梯度消失[15]。LSTM 层由一组循环连接的内存块组成，每一个内存块包含一个或多个循环连接的存储单元和三个乘法单元——输入门、输出门和遗忘门——为存储单元提供连续的写、读和重置操作[16]。

LSTM 单元结构如图 1 所示，其中 h_{t-1} 为过去隐藏状态，h_t 为当前隐藏状态，c_{t-1} 为过去细胞状态，c_t 为现在细胞状态，x_t 为当前输入信息，f_t 为遗忘门输出，i_t 为输入门输出，$\tilde{c}_t$ 为候选值向量，o_t 为输出门的输出。

具体流程为：将 h_{t-1} 和 x_t 一起传入遗忘门得到 f_t，通过 sigmoid 函数可将 f_t 的值控制在 0 和 1 之间，将接近 0 的值丢弃，接近 1 的值保留；将 h_{t-1} 和 x_t 同时传递到输入门的 sigmoid 和 tanh 函数中，分别得到 i_t 和 $\tilde{c}_t$，tanh 函数将 $\tilde{c}_t$ 的值控制在(−1, 1)，将i_t和 $\tilde{c}_t$ 相乘，由 i_t 决定 $\tilde{c}_t$ 丢弃或保留哪些信息；将 c_{t-1} 与 f_t 相乘的值与 i_t 和 $\tilde{c}_t$ 相乘的值相加，得到 c_t；将 h_{t-1} 和 x_t 一起传入输出门得到 o_t，将 c_t 传递给 tanh 函数，并将输出值与 o_t 相乘得到 h_t。

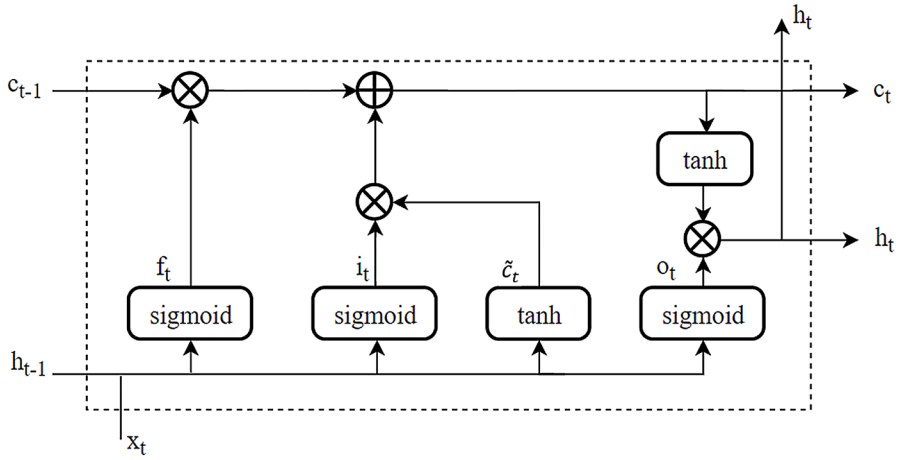

Figure 1. LSTM unit structure
图 1. LSTM 单元结构

LSTM 内存单元具体实现如下：

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

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

$$\tilde{c}_t = \tanh\left(W_c\left[h_{t-1}, x_t\right] + b_c\right)$$

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

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

$$h_t = o_t * \tanh\left(c_t\right)$$

其中 σ 是 sigmoid 函数，b 是偏置向量，W 是隐藏状态的权重矩阵。

LSTM 的隐藏状态只包含了过去的信息，但上下文信息对命名实体识别任务来说都十分重要。双向长短时记忆网络(BLSTM)由两个 LSTM 层组成，基本思想为使用两个 LSTM 层向前和向后形成独立的隐藏状态，通过正向的 LSTM 获得上文信息，逆向的 LSTM 获得下文的信息，然后将两个隐藏状态拼接起来作为最终输出[13]，能够高效完成序列标记任务。

3.2. CRF

BLSTM 模型只能够预测文本序列与标签的关系，而不能预测标签与标签之间的关系，其原理是输出概率最大值，当输出结果中概率值都很大，就可能会导致输出的两个标签顺序不合理。例如，使用 BIO 标签策略进行命名实体识别时，正确的标签序列中标签 O 后面不会出现标签 I。

条件随机场(CRF)能通过相邻标签之间的关联性求得最优的预测序列，可以弥补 BLSTM 无法处理相邻标签依赖关系的缺点[17] 。

在命名实体识别任务中，将 BLSTM 输出的标签序列输入到 CRF 中，使用 CRF 中的特征函数对标签序列进行打分，特征函数以当前词和其左边的词的标签为打分标准。打分函数如下：

$$score\left(L/S\right) = \sum_{j=1}^{m}\sum_{i=1}^{n}\lambda_j f_j\left(S, i, l_i, l_{i-1}\right)$$

其中 S 是需标记的句子，L 是标签序列，f 为特征函数，λ 为其权重，i 是词在句中的位置，l_i 是标签序列中第 i 个词的标签，m 为特征函数个数，n 为句子长度。将分数归一化为取值为 0 到 1 的概率值，公式如下：

$$P\left(L/S\right) = \frac{\exp\left[score\left(L/S\right)\right]}{\sum_{L'}\exp\left[score\left(L'/S\right)\right]}$$

计算所得的最大的概率值记为最优标签序列。

目前，BLSTM-CRF 是神经网络模型中使用频率最高的架构，BLSTM-CRF 模型如图 2 所示，其中灰色框表示 LSTM 单元。在 BLSTM-CRF 模型中，利用 BLSTM 来考虑上下文信息，进行高维特征抽取，同时利用 CRF 求得全局最优解。它可以使用过去的输入特征和句子级标签信息，以及未来的输入特征来预测当前标签。

3.3. BERT 模型

在众多语言模型中，Word2Vec 使用最广泛，但它学习到的语义信息受限于窗口大小[15]，而后 Peters 等人[18] 提出的 ELMo (Embeddings from Language Models)模型通过 BLSTM 对上下文信息进行建模，但

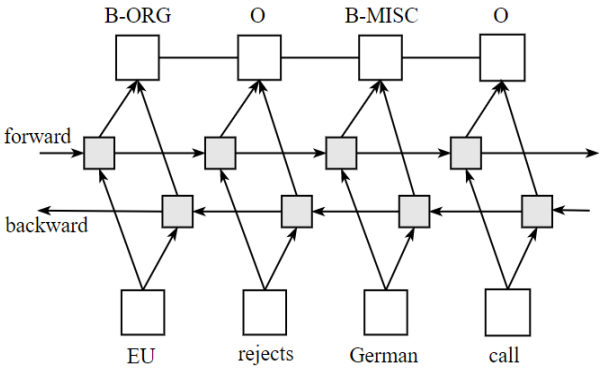

Figure 2. BLSTM-CRF model
图 2. BLSTM-CRF 模型

BLSTM 的信息抽取能力弱于 Transformer，同时，BLSTM 的序列特性使其无法进行并行计算。2018 年 Devlin 等人[19]提出的 BERT (Bidirectional Encoder Representations from Transformers)模型，使用多层 Transformer，能够同时获取句子前后两个方向上的信息，在诸多语言模型中取得了更好的成绩。

BERT 模型结构如图 3 所示，它使用了双向 Transformer 编码器结构，其中编码器框架使用层叠结构，包含多头注意力机制和前馈神经网络[20]。

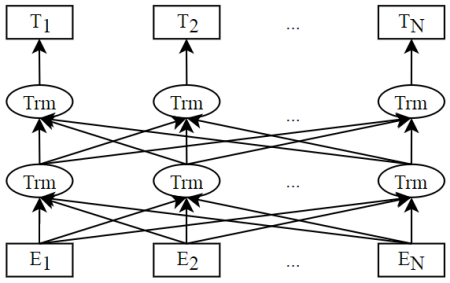

Figure 3. BERT model structure
图 3. BERT 模型结构

Transformer 的关键部分是注意力(attention)机制，它通过一个句子中的词与词之间的关联程度调整权重系数矩阵来获取词的表征[21]，得到一个有重要性程度区分的输出。Attention 机制的计算公式如下：

$$Attention(Q,K,V) = soft\max\left(\frac{QK^t}{\sqrt{d_k}}\right)V$$

其中 d_k 是输入向量维度，K、Q、V 是输入字向量矩阵。

Transformer 模型使用的多头注意力机制可以提高模型在不同位置的注意力单元的不同表示子空间，最终结果是将所有注意力单元的结果整合到一起。公式如下：

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \cdots, head_h)W^O$$

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

BERT 模型通过两个预训练任务“Mask 语言模型”和“下一句预测”分别获取词级别和句子级别的表示。在实际操作中，Mask 语言模型随机遮盖 15%的词，然后使用编码器预测被遮盖的词。下一句预测任务随机替换一些句子，利用上句预测下句。

3.4. BERT-BLSTM-CRF 模型

本文采用的 BERT-BLSTM-CRF 模型如图 4 所示。

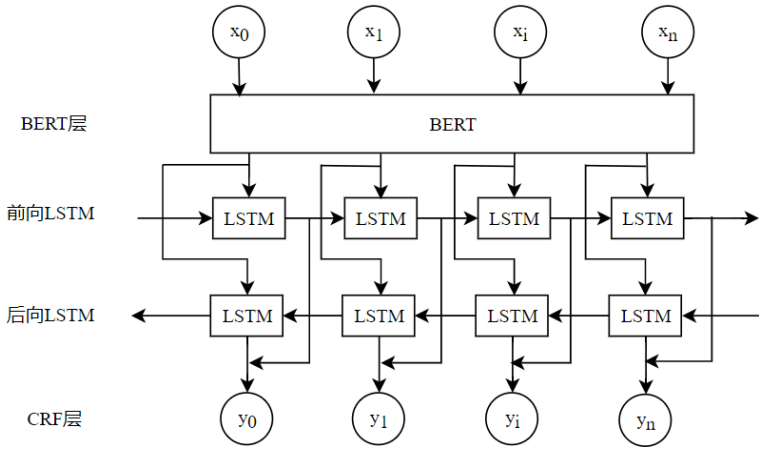

Figure 4. BERT-BLSTM-CRF model
图 4. BERT-BLSTM-CRF 模型

该模型的基本流程为：将标注语料通过 BERT 层得到融合了全文语义信息后的向量表示；然后将输出的向量输入到 BLSTM 层进行特征表示，通过前向 LSTM 获得当前词及其上文的向量，后向 LSTM 获得当前词及其下文的向量；最后将 BLSTM 层输出的各个标签概率输入到 CRF 层求得全局最优解。

4. 实验

4.1. 实验数据

本文实验使用北京大学计算语言研究所发布的人民日报语料库，这是目前国内构建的最大规模中文标注语料库，语料库中已经标注了人名、地名、机构名等信息。本文采用 BIO 标注，B 表示实体开始，I 表示实体中间词，O 表示无关词。将人名记为 PER，机构名记为 ORG，地名记为 LOC。例如，B-ORG 表示机构名实体的开始，I-PER 表示人名实体的中间词。实验过程中随机划分训练集、测试集和评估集，本实验中，训练集包含 20,864 个句子，6,277,429 个字；测试集 4636 个句子，1,405,788 个字；评估集包含 2318 个句子，702,455 个字。

4.2. 评价指标

本文采用准确率 P、召回率 R 和 F1 值作为模型性能的评价标准。具体定义如下：

$$P = \frac{\text{识别出的正确实体个数}}{\text{识别出的所有实体个数}} \times 100\%$$

$$R = \frac{\text{识别出的正确实体个数}}{\text{所有标注的实体个数}} \times 100\%$$

$$F_1 = \frac{2PR}{P+R} \times 100\%$$

4.3. 模型参数

本文使用了 Google 提供的 Bert，Chinese 模型，该模型使用 12 层 transformer，隐藏层为 768 层，12

头自注意力机制。其他实验参数如表 1 所示：

Table 1. Experimental parameters
表 1. 实验参数

参数	取值
seq_len	128
epochs	20
batch_size	64
Layer_dropout	0.4
Layer`_blstm	128

4.4. 实验结果

本文使用 BLSTM、BLSTM-CRF 和 BERT-BLSTM-CRF 模型对数据集进行训练，实验结果如表 2、表 3 所示。

Table 2. Recognition results of different entities by different models
表 2. 不同模型对不同实体的识别结果

模型名称	实体类型	准确率 P	召回率 R	F_1值
BLSTM	LOC	0.8236	0.8732	0.8477
	ORG	0.7332	0.7304	0.7318
	PER	0.9126	0.9313	0.9219
BLSTM-CRF	LOC	0.8596	0.8996	0.8792
	ORG	0.8498	0.7379	0.7899
	PER	0.9599	0.9459	0.9529
BERT-BLSTM-CRF	LOC	0.9304	0.9409	0.9356
	ORG	0.8773	0.8809	0.8791
	PER	0.9698	0.9687	0.9692

根据实验结果可以看出，地点和人名的识别效果比机构名的识别效果好，原因可能是在新闻文本中，对地名和人名的缩写和指代比较少。同时，从表中数据可以看出，BERT-BLSTM-CRF 模型对实体的识别结果与另外两种模型的识别结果相比有明显提升。

Table 3. Named entity recognition results of different models
表 3. 不同模型的命名实体识别结果

模型名称	准确率 P	召回率 R	F_1值
BLSTM	0.8197	0.8441	0.8317
BLSTM-CRF	0.8837	0.8615	0.8725
BERT-BLSTM-CRF	0.9245	0.9302	0.9274

通过表中结果可知，BLSTM-CRF 模型的性能优于 BLSTM 模型，说明 BLSTM 模型之后使用 CRF 对标签之间的关系进行考虑，可以提升整个模型的性能；通过 BLSTM-CRF 模型与 BERT-BLSTM-CRF

模型的对比，可以看出加入 BERT 预训练模型能更好地学习句子中词之间的关联性和重要程度，得到更好的词向量全局表达，进而提升模型性能。

5. 总结

本文对命名实体识别现状以及目前的研究进展进行了分析，针对中文实体识别的难点，将 BERT 模型与传统的 BLSTM-CRF 模型相结合，构造了 BERT-BLSTM-CRF 模型。该模型通过 BERT 模型计算句中词之间的关联性和每个词的权重来获得更全局的词向量表达，结合了 BLSTM 学习词语上下文信息的能力和 CRF 考虑全局信息推断标签的能力。通过在人民日报语料库上进行命名实体识别，并与其他传统模型相比，本文构造的 BERT-BLSTM-CRF 模型具有更好的性能。实验表明，将 BERT 模型运用于命名实体识别能起到提升性能的作用，对后续研究具有一定的参考价值。

参考文献

[1] Lample, G., Ballesteros, M., Subramanian, S., *et al*. (2016) Neural Architectures for Named Entity Recognition. *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, San Diego, 12-17 June 2016, 260-270. https://doi.org/10.18653/v1/N16-1030

[2] Yang, Z., Salakhutdinov, R. and Cohen, W. (2016) Multi-Task Cross-Lingual Sequence Tagging from Scratch.

[3] de Oliveira, D.M., Laender, A.H.F., Veloso, A., *et al*. (2013) FS-NER: A Lightweight Filter-Stream Approach to Named Entity Recognition on Twitter Data. *Proceedings of the* 22*nd International Conference on World Wide Web*, Rio de Janeiro, 13-17 May 2013, 597-604. https://doi.org/10.1145/2487788.2488003

[4] 罗熹, 夏先运, 安莹, 陈先来. 结合多头自注意力机制与 BiLSTM-CRF 的中文临床实体识别[J]. 湖南大学学报(自然科学版), 2021, 48(4): 45-55.

[5] 王昊, 林克柔, 孟镇, 李心蕾. 文本表示及其特征生成对法律判决书中多类型实体识别的影响分析[J/OL]. 数据分析与知识发现, 1-26. http://kns.cnki.net/kcms/detail/10.1478.G2.20210326.0959.002.html, 2021-04-25.

[6] 钟诗胜, 陈曦, 赵明航, 张永健. 引入词集级注意力机制的中文命名实体识别方法[J/OL]. 吉林大学学报(工学版), 1-7. https://doi.org/10.13229/j.cnki.jdxbgxb20200984, 2021-04-26.

[7] 何玉洁, 杜方, 史英杰, 宋丽娟. 基于深度学习的命名实体识别研究综述[J/OL]. 计算机工程与应用, 1-17. http://kns.cnki.net/kcms/detail/11.2127.TP.20210326.0937.002.html, 2021-04-21.

[8] 王玥. 基于深度学习的命名实体识别研究[D]: [硕士学位论文]. 昆明: 云南财经大学, 2019.

[9] 朱岩, 张利, 王煜. 基于 RoBERTa-WWM 的中文电子病历命名实体识别[J]. 计算机与现代化, 2021(2): 51-55.

[10] Collobert, R., Weston, J., Bottou, L., *et al*. (2011) Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research*, **12**, 2493-2537.

[11] Huang, Z., Xu, W. and Yu, K. (2015) Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv preprint arXiv:1508.01991.

[12] Chiu, J.P.C. and Nichols, E. (2015) 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

[13] 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, 7-12 August 2016, 1064-1074. https://doi.org/10.18653/v1/P16-1101

[14] Rei, M., Crichton, G.K.O. and Pyysalo, S. (2016) Attending to Characters in Neural Sequence Labeling Models. *Proceedings of COLING* 2016, *the* 26*th International Conference on Computational Linguistics*: *Technical Papers*, Osaka, 11-16 December 2016, 309-318.

[15] 王子牛, 姜猛, 高建瓴, 陈娅先. 基于 BERT 的中文命名实体识别方法[J]. 计算机科学, 2019, 46(S2): 138-142.

[16] Graves, A. and Schmidhuber, J. (2005) Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures. *Neural Networks*, **18**, 602-610. https://doi.org/10.1016/j.neunet.2005.06.042

[17] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 4-9 December 2017, 5998-6008.

[18] Peters, M.E., Neumann, M., Iyyer, M., *et al*. (2018) Deep Contextualized Word Representations. *Proceedings of NAACL-HLT* 2018, New Orleans, 1-6 June 2018, 2227-2237.

[19] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT* 2019, Minneapolis, 2-7 June 2019, 4171-4186.

[20] 袁培森, 李润隆, 王翀, 徐焕良. 基于 BERT 的水稻表型知识图谱中关系抽取研究[J/OL]. 农业机械学报, 1-10. http://kns.cnki.net/kcms/detail/11.1964.S.20210315.1809.015.html, 2021-04-26.

[21] 谢腾, 杨俊安, 刘辉. 基于 BERT-BiLSTM-CRF 模型的中文实体识别[J]. 计算机系统应用, 2020, 29(7): 48-55.