

# 融合词信息与自注意力的教育命名实体识别

郑守民, 申艳光

(河北工程大学信息与电气工程学院, 河北 邯郸 056004)

**摘要:** 为解决教育领域命名实体识别任务精度较低和语料严重不足的问题, 提出一种融合词信息与自注意力的命名实体识别模型 WBBAC。该模型利用 BERT 预训练语言模型增强字向量的语义表示并为字向量引入词频信息, 将字向量与词向量拼接作为双向长短期记忆网络的输入, 经过自注意力层进一步寻找序列内部的联系, 最后通过 CRF 解码获得最优序列。根据课程文本特点创建计算机组成原理数据集并进行标注, 在 Resume 数据集和计算机组成原理数据集上进行实验, WBBAC 模型的 F1 值分别为 95.65% 和 73.94%。实验结果表明, 与基线模型相比, WBBAC 模型具有更高的 F1 值, 有效解决了教育领域命名实体识别任务中标注数据不足的问题。

**关键词:** 命名实体识别; 词信息; 自注意力机制; 教育领域; BERT

DOI: 10.11907/rjdk.231921

开放科学(资源服务)标识码(OSID):



中图分类号: TP391

文献标识码: A

文章编号: 1672-7800(2024)009-0105-05

## Educational Named Entity Recognition Integrating Word Information and Self-Attention Mechanism

ZHENG Shoumin, SHEN Yanguang

(School of Information and Electrical Engineering, Hebei University of Technology, Handan 056004, China)

**Abstract:** To solve the problems of low accuracy and severe lack of corpus in named entity recognition tasks in the education field, a named entity recognition model WBBAC that integrates word information and self attention is proposed. This model utilizes a BERT pre trained language model to enhance the semantic representation of word vectors and introduces word frequency information into them. The word vectors are concatenated with each other as inputs to a bidirectional long short-term memory network, which further searches for internal connections within the sequence through a self attention layer. Finally, the optimal sequence is obtained through CRF decoding. Create a computer composition principle dataset based on the characteristics of the course text and annotate it. Conduct experiments on the Resume dataset and the computer composition principle dataset, and the F1 values of the WBBAC model are 95.65% and 73.94%, respectively. The experimental results show that compared with the baseline model, the WBBAC model has a higher F1 value, effectively solving the problem of insufficient annotated data in named entity recognition tasks in the education field.

**Key Words:** named entity recognition; word information; self-attention mechanism; education domain; BERT

## 0 引言

随着互联网技术的快速发展, 全球范围内的线上教育迅速普及<sup>[1]</sup>。大学学科中涵盖着广泛的知识点, 课程内容在书籍中通常以目录章节的形式呈现, 然而这些章节所囊括的细小知识点及其相互之间的关联却不能一目了然, 可能会对学生的学习效果造成影响。利用知识图谱可以有

效解决这一问题, 而命名实体识别(Named Entity Recognition, NER)是知识图谱构建过程中非常关键的一个步骤。NER是自然语言处理(Natural Language Processing, NLP)中一项基本任务, 旨在从非结构化文本中识别出特定意义和类型的实体<sup>[2]</sup>。早期的NER方法使用基于规则的系统, 依赖于手工制作的语言规则和词典, 无法应对自然语言的多样性和复杂性<sup>[3]</sup>。近年来, 基于统计机器学习的方法开始流行起来, 如隐马尔可夫模型<sup>[4]</sup>、条件随机场(Conditional

收稿日期: 2023-08-28

扫描二维码阅读全文:



**作者简介:** 郑守民(1997-), 男, CCF会员, 河北工程大学信息与电气工程学院硕士研究生, 研究方向为智能处理及应用; 申艳光(1970-), 女, 河北工程大学信息与电气工程学院教授、硕士生导师, 研究方向为机器学习、数据挖掘。本文通讯作者: 申艳光。

Random Fields, CRF)<sup>[5]</sup>、支持向量机<sup>[6]</sup>等。但统计机器学习方法需要人工选择和设计特征,耗费大量人力,泛化能力也十分有限。相较于医学、金融等比较热门的领域而言,教育领域对于NER的研究和实践较少;同时教育命名实体专业术语较多,制约了NER技术在教育领域的广泛应用。

### 1 相关研究

近年来,基于深度学习的NER技术得到快速发展,在特征提取深度和模型精确度方面已经逐渐超过传统基于机器学习的方法,成为目前的主流方法<sup>[7]</sup>。例如,文献<sup>[8]</sup>将长短期记忆网络(Long Short-Term Memory, LSTM)应用于命名实体识别任务;文献<sup>[9]</sup>提出BiLSTM-CRF模型,利用双向长短期记忆网络(Bi-directional Long Short-Term Memory, BiLSTM)提取上下文特征,结合CRF学习数据集中不同标签之间的转换概率,从而对BiLSTM层的输出进行修正;文献<sup>[10]</sup>在BiLSTM-CRF模型中加入自注意力机制,增强了对上下文信息的捕捉,提升了模型识别精度。然而以上方法在处理教育领域文本时得到的结果并不理想。此外,中文NER相较于英文NER难度更大,因为英文有着更加明确的词边界<sup>[11]</sup>。同时教育领域文本中涉及的实体专业性较强,中文分词易出现溢出词(Out-Of-Vocabulary, OOV)问题,从而使NER效果变差。文献<sup>[12]</sup>研究表明,采用基于字符的方法可以避免中文NER中的分词错误。

在NLP领域,为使词语能包含丰富的语义信息及句法特征,研究人员陆续提出使用预训练语言模型获取词语的表示,代表性模型包括ELMo<sup>[13]</sup>、GPT<sup>[14]</sup>和BERT(Bidirectional Encoder Representations from Transformers)<sup>[15]</sup>。然而,基于字符的方法没有充分利用数据中的单词信息,而这些单词信息对于识别专业领域的实体非常重要。为此文献<sup>[16]</sup>提出Lattice LSTM模型,使得词信息得到了充分的利用,在中文NER上取得了很好的效果,但该模型训练速度非常慢。

目前,针对通用领域的NER方法已经取得了一定成果<sup>[17]</sup>。然而,在专业领域中,由于缺乏相应的标注数据,实体识别变得十分困难。因此,近年来,利用深度学习方法解决专业领域的NER问题成为研究热点。在教育领域,实体识别面临着与通用领域不同的挑战,不仅是课程相关文本中涉及的术语具有很强的专业性,而且存在着实体之间相互包含嵌套的情况,导致实体构成非常复杂。传统NER方法在处理这些实体时往往存在识别正确率较低和覆盖率不足的问题。为此,本文提出一种融合词信息与自注意力的模型WBBAC(Word BERT BiLSTM Attention CRF)应用于教育领域。该模型首先将词向量和字向量拼接送入BiLSTM,通过统计词频的方式降低命名实体词边界划分错误的影响,同时使用BERT预训练模型获取更丰富的语义特征;然后利用自注意力机制加强对有效信息的关注而忽略次要信息,从而获得更好的识别效果。针对教育领域

已有数据集缺乏的问题构建了一个计算机组成原理课程的NER数据集,以考察所提模型的有效性,同时在Resume公共数据集上评价该模型的泛化能力。

### 2 WBBAC模型

WBBAC模型由词嵌入层、BiLSTM层、自注意力层和CRF层4个模块组成。总体结构如图1所示。在WBBAC模型中,首先使用BERT预训练语言模型生成对应文本的字向量;然后将字向量与使用Word2Vec训练生成的词向量拼接到一起送入BiLSTM层进行上下文信息的特征提取,再经过自注意力层对从BiLSTM层得到的输出进行权重调节,以加强对重要信息的关注;最后通过CRF层进行标签推理,从而获得最优序列。

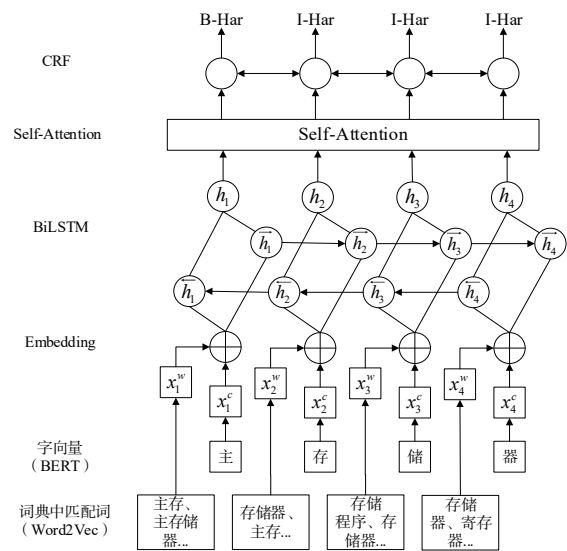


Fig. 1 Overall structure of model

图1 模型总体结构

#### 2.1 词嵌入层

通过词嵌入层将经过BERT预训练语言模型生成的字向量与使用Word2Vec训练得到的词向量融合在一起,充分利用了文本中的词信息,而不是使用单一的字向量作为输入。

##### 2.1.1 基于BERT的字向量表示

BERT是一种基于Transformer架构的预训练语言模型,其能够理解文本中的上下文信息。教育领域的知识点实体之间往往具有复杂的关系,BERT可以捕捉到这些实体在不同上下文中的含义,从而提高NER任务的准确性。BERT的输入是经过特殊处理的文本序列,这些处理确保了模型能从上下文中获取丰富的语义信息。BERT的输入向量表示由Token Embeddings、Segment Embeddings和Positional Embeddings 3个部分组成,具体如图2所示。其中,Token Embedding是对输入文本中的每个token进行嵌入映射得到的向量表示;Segment Embeddings为片段信息,用于指示每个token属于哪个句子;Positional Embeddings为位置信息,用于指示模型每个token在序列中的位置。

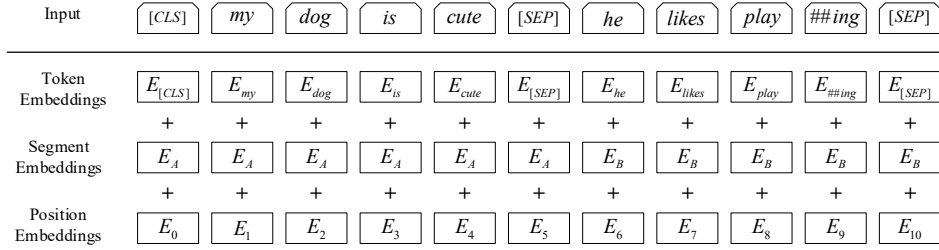


Fig. 2 BERT input vector representation

图2 BERT输入向量表示

2.1.2 词向量表示

传统的中文NER模型往往需要先进行中文分词,分词效果的好坏直接影响识别结果,例如分词错误可能会导致错误的词汇边界,进而引发误差传播问题。为此,本文将句子与词典进行匹配,对每个字符对应的词语进行词向量表示,通过统计词频的方式为错误分词分配一个较低的权重,从而减少分词错误对模型性能的影响。同时,将模型引入词信息的方法也解决了在NER任务中可能遇到的数据稀疏性和词汇不平衡问题,因为该方法为每个字符分配了初始词汇词频权重,可以为模型提供一些关于词汇的初步语义信息。

对于序列中的每个字符,通过归一化词频对字符所对应词汇的词向量进行加权,对于不包含匹配词语的字符则使用None补齐。以词频为权重考虑每个词语的重要性,从而更好地提取语义特征。计算公式为:

$$e_i^w = \begin{cases} \sum_{w \in S} z(w)e^w(w) / \sum_{w \in S} z^w(w), S = \emptyset \\ e^w( None ), S = \emptyset \end{cases} \quad (1)$$

式中: $S$ 表示字符对应词语的合集; $z(w)$ 表示词 $w$ 的词频,通过对数据集文本进行统计得出; $e^w(w)$ 表示词 $w$ 的词向量,通过Word2Vec在大量领域文本上训练得到; $e^w( None )$ 表示None的词向量。

词典通过对大量领域文本进行分词后再经过一定的人工筛选得到,可最大程度地避免分词错误带来的影响。本文选择的分词工具为JieBa。将字向量与词向量拼接得到该字符的向量表示。公式为:

$$e_i = e_i^w \oplus e_{BERT}^c(c_i) \quad (2)$$

2.2 BiLSTM层

LSTM是一种循环神经网络架构,旨在处理序列数据并克服传统循环神经网络序列长度过长时会出现的梯度消失问题。LSTM单元结构如图3所示。相较于传统的循环神经网络,LSTM神经元结构采用3个门控单元,分别为遗忘门、输入门和输出门,通过保持和更新网络状态实现长期记忆功能。

LSTM的参数化表示为:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

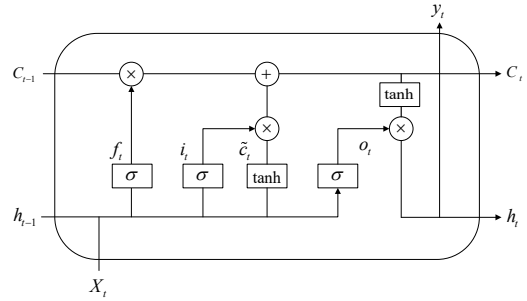


Fig. 3 LSTM unit structure

图3 LSTM单元结构

$$\tilde{c}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (6)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (7)$$

$$h_t = o_t \odot \tanh(c_t) \quad (8)$$

式中: $\sigma$ 表示sigmoid激活函数; $x_t$ 表示 $t$ 时刻的输入向量; $\odot$ 表示点乘运算; $i_t, f_t$ 和 $o_t$ 分别表示 $t$ 时刻的输入门、遗忘门和输出门; $W$ 和 $b$ 对应每个门的权重矩阵和偏置向量。

BiLSTM是一种变种的LSTM网络,由前向LSTM和后向LSTM组成,能够更好地提取序列中的上下文特征。BiLSTM的隐藏状态由前向和后向LSTM的隐藏状态在每个时刻连接在一起得到。表示为:

$$h_t = [h_t^f, h_t^b] \quad (9)$$

式中: $h_t^f$ 和 $h_t^b$ 分别表示 $t$ 时刻前向和后向LSTM的隐藏状态。

2.3 自注意力层

自注意力机制允许模型中每个在输入序列中的词都能够关注到其他所有词汇的信息,而不仅仅是像BiLSTM那样依赖于过去和未来的隐藏状态。将自注意力层与BiLSTM结合可以为教育领域的NER任务提供更好的上下文建模、长距离依赖捕获以及权重分配能力,从而提高模型性能。

通过BiLSTM层对文本进行特征提取后得到输出值 $h$ ,将 $h$ 与3个权重矩阵 $W_q, W_k, W_v$ 相乘得到 $Q, K, V$ 3个向量。在计算某个单词的权重时,将 $Q$ 与 $K$ 进行相似度计算,得出权重后,将权重系数与 $V$ 进行加权求和,得到最终结果 $Z$ 。相关计算公式为:

$$Attention(Q, K, V) = \text{soft max} \left( \frac{QK^T}{\sqrt{d_k}} \right) V \quad (10)$$

式中: $d_k$ 表示 $Q$ 、 $K$ 、 $V$ 的维度。本文使用的多头自注意力机制本质上是进行多次自注意力机制计算,然后将多次自注意力机制结果进行拼接,从而进行更深层次的特征提取。计算公式为:

$$head_i = Attention(QW_i^q, KW_i^k, VW_i^v) \quad (11)$$

$$Multi(Q, K, V) = Concat(head_1, \dots, head_h)W^o \quad (12)$$

## 2.4 CRF推理层

CRF是用于处理NER的常见机器学习算法,其能够从训练数据中学习到约束条件,以确保预测标签的有效性。输入序列 $X = (x_1, x_2, \dots, x_n)$ 及相对应的输出标签序列 $Y = (y_1, y_2, \dots, y_n)$ 的评估分数计算公式为:

$$score(x, y) = \sum_{i=0}^n A_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i} \quad (13)$$

式中: $A$ 表示转移得分矩阵; $A_{ij}$ 表示从标签 $i$ 到标签 $i+1$ 的转移分数; $P$ 表示输出得分矩阵; $P_{ij}$ 表示第 $i$ 个词在第 $j$ 个标签下的输出得分; $n$ 表示序列长度。

使用softmax函数归一化得到预测序列的概率。计算公式为:

$$P(y|x) = \frac{e^{s(x, y)}}{\sum_{\tilde{y} \in \tilde{Y}_x} e^{s(x, \tilde{y})}} \quad (14)$$

式中: $\tilde{y}$ 表示真正标记值, $\tilde{y}_x$ 为所有可能的标记序列。

训练过程中最大化预测序列的概率。计算公式为:

$$\log_a P(y|x) = s(x, y) - \sum_{\tilde{y} \in \tilde{Y}_x} s(x, \tilde{y}) \quad (15)$$

最后利用Viterbi算法得到文本序列标签的最优序列。

## 3 实验方法与结果分析

### 3.1 实验环境

在Python3.6, Pytorch1.8.0环境下进行实验。预训练语言模型BERT采用Google预训练模型BERT-base版本,隐藏层维度为768,优化器采用Adam算法。实验参数设置如表1所示。

Table 1 Parameter setting

表1 参数设置

参数	含义	值
max_len	最大序列长度	128
dropout	丢失率	0.5
batch_size	数据批次大小	32
lr	学习率	0.001
heads	自注意力头数	12
wordvec_dom	词向量维度	100
epoch	迭代次数	100

### 3.2 数据集

本实验使用两组数据集,一组是Resume公共数据集,另一组是自建计算机组成原理数据集。自建数据集主要通过爬虫获得,抓取百度百科有关计算机组成原理课程知

识点的文本语料,并通过PCR技术对文献[18]进行扫描以获取原始文本语料,最终得到2312个计算机组成原理知识点相关句子。根据计算机组成原理课程的特点将知识点实体划分为3类,具体如表2所示。

Table 2 Entity types statistics

表2 实体类型信息

类别符号	实体类别	实体描述
HAR	硬件知识点	课程中描述硬件的知识点,如主存储器
SOF	软件知识点	课程中描述软件的知识点,如汇编语言
REL	相关术语	非以上两种类型的知识点,如无符号数

对获取到的原始文本语料进行清洗和去重,使用开源软件YEDDA进行人工标注。自建数据集采用BIO标注,其中B(Begin)表示一个实体的开始部分,I(Inside)表示实体内部,O(Outside)表示不属于实体的部分,在实体后还要加上类别信息,例如“运B-HAR/算I-HAR/器I-HAR/用O/来O/完O/成O/算B-REL/术I-REL/运I-REL/算I-REL/和O/逻B-REL/辑I-REL/运I-REL/算I-REL”。

按照8:1:1的比例将数据集划分为训练集、验证集和测试集。两个数据集具体信息如表3所示。

Table 3 Statistics of datasets

表3 数据集统计信息

数据集	类型	×10 <sup>3</sup>		
		训练集	验证集	测试集
Resume	句子	3.8	0.5	0.5
	字符	124.1	13.9	15.1
计算机组成原理	句子	2.3	0.3	0.3
	字符	75.2	9.7	10.1

### 3.3 评价指标

使用精确率 $P$ (Precision)、召回率 $R$ (Recall)和 $F1$ 值( $F1$ -Score)作为模型评价指标。其中, $P$ 表示正确识别的实体数与总被识别的实体数之比; $R$ 表示正确识别的实体数与文本中所有实体数之比; $F1$ 表示精确率与召回率的调和平均值。

$$P = \frac{T_p}{T_p + F_p} \times 100\% \quad (16)$$

$$R = \frac{T_p}{T_p + F_N} \times 100\% \quad (17)$$

$$F1 = 2 \times \frac{P \times R}{P + R} \times 100\% \quad (18)$$

式中: $T_p$ 表示正确识别的实体数; $F_p$ 表示识别到与正确实体不相关的实体数; $F_N$ 表示错误识别的实体数。

### 3.4 实验结果

为验证本文模型的有效性,选择BERT-BiLSTM-CRF模型<sup>[19]</sup>作为Baseline进行比较实验。为了验证添加模块的有效性,分别在Baseline上只添加词信息和自注意力机制<sup>[20]</sup>进行比较试验。为验证BERT预训练模型的有效性,选择BiLSTM-CRF模型以及同样引入词汇信息的Lattice-LSTM模型进行比较试验。实验结果如表4所示。可以看出,BiLSTM-CRF模型和融合了词汇信息的Lattice-LSTM模型的 $F1$ 值明显低于其他4个模型,表明预训练模型

Table 4 Experimental results of different models

模型	表4 不同模型实验结果比较						%
	Resume数据集			计算机组成原理数据集			
	P	R	F1	P	R	F1	
BiLSTM-CRF	93.66	93.31	93.48	69.32	69.08	69.20	
Lattice-LSTM	94.81	94.11	94.46	71.38	70.25	70.81	
BERT-BiLSTM-CRF(Baseline)	95.10	95.23	95.16	73.28	72.56	72.92	
Baseline+Attention	95.67	95.28	95.47	73.93	73.22	73.57	
Baseline+word	95.22	95.60	95.41	73.65	73.39	73.52	
WBBAC(本文模型)	95.79	95.52	95.65	74.11	73.78	73.94	

BERT对于识别效果的提升相当大。在Baseline基础上单独添加词信息的模型在Resume和计算机组成原理两个数据集上的F1值相较Baseline均有一定提升,表明词信息的加入对于识别效果提升有一定作用。单独添加自注意力机制的模型F1值高于单独添加词信息的模型,但是召回率低于单独添加词信息的模型,表明自注意力机制与词信息两个模块对于识别效果提升均有效,且自注意力机制的作用更大,而词频信息的加入使模型更倾向于查全。所有模型在Resume数据集上的F1值远高于计算机组成原理数据集的F1值,这是由于自建数据集的标注质量不如Resume公共数据集。融合了词信息和自注意力机制的本文模型在Resume和计算机组成原理两个数据集中的F1值分别为95.65%和73.94%,均取得了最好效果,有效性和泛化性得到验证。

## 4 结语

本文针对教育领域NER精确率较低的问题提出一种融合词信息与自注意力机制的中文NER模型。该模型将词信息融合进基于字符的模型中,使词信息得到充分利用,同时采用自注意力机制增强对关键信息的关注,寻找序列内部联系,在公共数据集和自建数据集中均获得了较好的识别效果。然而,本文模型在Resume数据集上的P、R、F1值远高于自建课程数据集,这是由于自建数据集的标注质量不如Resume公共数据集,且计算机组成原理课程文本专业性较强。因此,提高自建数据集的质量将成为后续重点研究方向之一。

### 参考文献:

- [1] TU J H, XIAO J Y, JIANG G F. Building a calculus knowledge graph to boost first-class course development[J]. China University Teaching, 2020(11): 33-37.  
涂建华,肖珺怡,姜广峰. 构建微积分知识图谱助推一流课程建设[J]. 中国大学教学, 2020(11): 33-37.
- [2] LIU L, WANG D B. Summary of research on named entity recognition[J]. Acta Information, 2018, 37(3): 329-340.  
刘浏,王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329-340.
- [3] EFTIMOV T, SELJAK B K, KOROSIC P. A rule-based named entity recognition method for knowledge extraction of evidence-based dietary recom-

- mendations[J]. PLoS ONE, 2017, 12(6): e0179488.
- [4] TODOROVIC B T, RANCIC S R, MARKOVIC I M, et al. Named entity recognition and classification using context hidden Markov model[C]//Proceedings of 2008 9th Symposium on Neural Network Applications in Electrical Engineering, 2008: 43-46.
- [5] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning, 2001: 282-289.
- [6] JU Z, WANG J, ZHU F. Named entity recognition from biomedical text using SVM[C]//5th International Conference on Bioinformatics and Biomedical Engineering, 2011: 1-4.
- [7] HE Y J, DU F, SHI Y J, et al. Survey of named entity recognition based on deep learning[J]. Computer Engineering and Applications, 2021, 57(11): 21-36.  
何玉洁,杜方,史英杰,等. 基于深度学习的命名实体识别研究综述[J]. 计算机工程与应用, 2021, 57(11): 21-36.
- [8] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [9] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[DB/OL]. <https://arxiv.org/abs/1508.01991>.
- [10] LUO L, YANG Z, YANG P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition[J]. Bioinformatics, 2018, 34(8): 1381-1388.
- [11] MAHANTA H. A study on the approaches of developing a named entity recognition tool[J]. International Journal of Research in Engineering and Technology, 2013, 2(14): 58-61.
- [12] HE J, WANG H. Chinese named entity recognition and word segmentation based on character[C]//Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing, 2008: 128-132.
- [13] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representation[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, 2018: 2227-2237.
- [14] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pretraining[EB/OL]. [https://blog.csdn.net/leo\\_95/artic-le/detail/89965558](https://blog.csdn.net/leo_95/artic-le/detail/89965558).
- [15] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, 2019: 4171-4186.
- [16] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1554-1564.
- [17] PENG M L, XING X Y, ZHANG Q, et al. Distantly supervised named entity recognition using positive-unlabeled learning[DB/OL]. <http://arxiv.org/abs/1906.01278>.
- [18] TANG S F. Principles of computer composition[M]. Beijing: Higher Education Press, 2020.  
唐朔飞. 计算机组成原理[M]. 北京: 高等教育出版社, 2020.
- [19] XIE T, YANG J A, LIU H. Chinese entity recognition based on BERT-BiLSTM-CRF Model[J]. Computer Systems & Applications, 2020, 29(7): 48-55.  
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.
- [20] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000-6010.