计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(7):48−55 [doi: 10.15888/j.cnki.csa.007525]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 BERT-BiLSTM-CRF 模型的中文实体识别①

谢　腾, 杨俊安, 刘　辉

(国防科技大学 电子对抗学院, 合肥 230037)
通讯作者: 杨俊安, E-mail: yangjunan@ustc.edu

摘　要: 命名实体识别是自然语言处理的一项关键技术. 基于深度学习的方法已被广泛应用到中文实体识别研究中. 大多数深度学习模型的预处理主要注重词和字符的特征抽取, 却忽略词上下文的语义信息, 使其无法表征一词多义, 因而实体识别性能有待进一步提高. 为解决该问题, 本文提出了一种基于 BERT-BiLSTM-CRF 模型的研究方法. 首先通过 BERT 模型预处理生成基于上下文信息的词向量, 其次将训练出来的词向量输入 BiLSTM-CRF 模型做进一步训练处理. 实验结果表明, 该模型在 MSRA 语料和人民日报语料库上都达到相当不错的结果, *F*1 值分别为 94.65% 和 95.67%.

关键词: 命名实体识别; BERT 模型; 双向长短期记忆网络; 条件随机场; 词向量

引用格式: 谢腾,杨俊安,刘辉.基于 BERT-BiLSTM-CRF 模型的中文实体识别.计算机系统应用,2020,29(7):48–55. http://www.c-s-a.org.cn/1003-3254/7525.html

Chinese Entity Recognition Based on BERT-BiLSTM-CRF Model

XIE Teng, YANG Jun-An, LIU Hui

(College of Electronic Engineering, National University of Defense Technology, Hefei 230037, China)

Abstract: Named Entity Recognition is a key technology in natural language processing, and the methods based on deep learning have been widely used in Chinese entity recognition. Most deep learning models focus on the feature extraction of words and characters, but ignore the semantic information of word context, therefore, they cannot represent polysemy, and the performance of entity recognition needs to be further improved. In order to solve this problem, this study proposes a method based on the BERT-BiLSTM-CRF model. First, word vectors based on context information are generated by the pretreatment of BERT model, and then the trained word vector is input into BiLSTM-CRF model for further training. The experimental result shows that the proposed model achieves sound results and reaches *F*1-score of 94.65% and 95.67% respectively in the MSRA corpus and People's Daily.

Key words: Named Entity Recognition (NER); BERT model; BiLSTM; CRF; word vector

引言

命名实体识别 (Named Entity Recognition, NER) 是自然语言处理的关键技术之一, 同时也是作为知识抽取的一项子任务, 其主要作用就是从海量文本中识别出特定类别的实体, 例如人名、地名、组织机构名以及领域专有词汇等. 中文命名实体识别是信息抽取、信息检索、知识图谱、机器翻译和问答系统等多种自然语言处理技术必不可少的组成部分, 在自然语言处理技术走向实用化的过程中占有重要地位. 因此, 命名实体识别作为自然语言处理最基础的任务, 对它的研究则具有非凡的意义与作用. 在中文实体识别任务中, 其难点主要表现在以下几个方面: (1) 命名实体

① 基金项目: 安徽省自然科学基金 (1908085MF202); 国防科技大学校基金 (ZK18-03-14)
Foundation item: Natural Science Foundation of Anhui Province (1908085MF202); Fund of National University of Defense Technology (ZK18-03-14)
收稿时间: 2019-12-22; 修改时间: 2020-01-19; 采用时间: 2020-02-11; csa 在线出版时间: 2020-07-03

类型与数量众多, 而且不断有新的实体涌现, 如新的人名、地名等; (2) 命名实体构成结构较复杂, 如组织机构存在大量的嵌套、别名以及缩略词等问题, 没有严格的命名规律; (3) 命名实体识别常常与中文分词、浅层语法分析等相结合, 而这两者的可靠性也直接决定命名实体识别的有效性, 使得中文命名实体识别更加困难. 因此, 中文命名实体识别研究还存在很大的提升空间, 有必要对其做进一步的研究.

1 相关工作

命名实体识别从最早期开始, 主要是基于词典与规则的方法, 它们依赖于语言学家的手工构造的规则模板, 容易产生错误, 不同领域间无法移植. 因此, 这种方法只能处理一些简单的文本数据, 对于复杂非结构化的数据却无能为力. 随后主要是基于统计机器学习的方法, 这些方法包括隐马尔可夫模型 (HMM)、最大熵模型 (MEM)、支持向量机 (SVM) 和条件随机场 (CRF) 等. 例如, 彭春艳等人[1]就利用 CRF 结合单词结构特性与距离依赖性, 在生物命名实体上取得较好的结果; 鞠久朋等人[2]提出把 CRF 与规则相结合来进行地理空间命名实体识别, 该算法有效地提高了地理空间命名实体识别的性能; 乐娟等人[3]提出基于 HMM 的京剧机构命名实体识别算法, 并且取得相当不错的效果. 在基于机器学习的方法中, NER 被当作序列标注问题, 利用大规模语料来学习标注模型. 但是这些方法在特征提取方面仍需要大量的人工参与, 且严重依赖于语料库, 识别效果并非很理想. 近些年来, 深度学习被应用到中文命名实体识别研究上. 基于深度学习的方法, 是通过获取数据的特征和分布式表示, 避免繁琐的人工特征抽取, 具有良好的泛化能力. 最早使用神经网络应用到命名实体研究上是 Hammerton 等人[4], 他们使用单向的长短期记忆网络 (LSTM), 该网络具有良好的序列建模能力, 因此 LSTM-CRF 成为了实体识别的基础架构; 后来在该模型的基础上, Guillaume Lample 等人[5]提出双向长短期记忆网络 (Bidirectional Long Short-Term Memory, BiLSTM) 和条件随机场 (CRF) 结合的神经网络模型, 这种双向结构能够获取上下文的序列信息, 因此在命名实体识别等任务中得到相当广泛的应用, 并且他们利用 BiLSTM-CRF 模型在语料库 CoNLL-2003 取得了比较高的 $F1$ 值 90.94%; Collobert 等人[6]就首次使用 CNN 与 CRF 结合的方式应用于命名实体识别研究中, 在 CoNLL-2003 上取得不错的效果; Huang 等人[7]在 BiLSTM-CRF 模型的基础上融入人工设计的拼写特征, 在 CoNLL-2003 语料上达到了 88.83% 的 $F1$ 值; Chiu 和 Nichols 等人[8]在 LSTM 模型前端加入 CNN 处理层, 在 CoNLL-2003 语料库上达到了 91.26% 的 $F1$ 值; 在生物医学领域上, 李丽双等人[9]利用 CNN-BiLSTM-CRF 神经网络模型在 Biocreative Ⅱ GM 和 JNLPBA2004 语料上取得了目前最好的 $F1$ 值, 分别为 89.09% 和 74.40%; 在化学领域上, Ling Luo 等人[10]采用基于 attention 机制的 BiLSTM-CRF 模型, 在 BioCreative IV 数据集上取得 91.14% 的 $F1$ 值; Fangzhao Wu 等人[11]提出联合分词与 CNN-BiLSTM-CRF 模型共同训练, 增强中文 NER 模型实体识别边界的能力, 同时又介绍了一种从现有标记数据中生成伪标记样本的方法, 进一步提高了实体识别的性能; 秦娅等人[12]在深度神经网络模型的基础上, 提出一种结合特征模板的 CNN-BiLSTM-CRF 网络安全实体识别方法,利用人工特征模板提取局部上下文特征, 在大规模网络安全数据集上 $F1$ 值达到 86%; 武惠等人[13]联合迁移学习和深度学习应用到中文 NER 上, 也取得了较好的效果; 王红斌[14]、王银瑞[15]利用迁移学习来进行实体识别, 该方法相对监督学习方法很大程度上减少了人工标注语料的工作量; Dong 等[16]提出了 Radical-BiLSTM-CRF 模型使用双向 LSTM 提取字根序列的特征, 然后与字向量拼接组成模型的输入; 刘晓俊等人[17]利用基于 attention 机制的 DC-BiLSTM-CRF 模型在 MSRA 语料上 $F1$ 值最高可达到 92.05%; Zhang 等人[18]提出的 Lattice LSTM 模型, 它显式地利用了词与词序列信息, 避免了分词错误的传递, 在 MSRA 语料上取得了较高的 $F1$ 值 93.18%; Liu 等人[19]提出 WC-LSTM 模型, 把词信息加入到整个字符的开头或末尾, 增强语义信息, 在 MSRA 语料上取得了 93.74% 的 $F1$ 值; 王蕾等人[20]则是利用片段神经网络结构, 实现特征的自动学习, 并在 MSRA 语料上取得 90.44% 的 $F1$ 值.

然而以上方法存在这样的一个问题: 这些方法无法表征一词多义, 因为它们主要注重词、字符或是词与词之间的特征提取, 而忽略了词上下文的语境或语义, 这样提取出来的只是一种不包含上下文语境信息的静态词向量, 因而导致其实体识别能力下降. 为解决该问题, 谷歌团队 Jacob Devlin 等人[21]所提出来一种

BERT (Bidirectional Encoder Representation from Transformers) 语言预处理模型来表征词向量, BERT 作为一种先进的预训练词向量模型, 它进一步增强词向量模型泛化能力, 充分描述字符级、词级、句子级甚至句间关系特征, 更好地表征不同语境中的句法与语义信息. Fábio Souza 等人[22]采用 BERT-CRF 模型应用到 Portuguese NER 上, 在 HAREM I 上取得最佳的 $F1$ 值; Jana Straková等人[23]把 BERT 预处理模型应用到实体识别上, 在 CoNLL-2002 Dutch、Spanish 和 CoNLL-2003 English 上取得相当理想的效果. 由于 BERT 具有表征一词多义的能力, 本文在此基础上提出一种 BERT-BiLSTM-CRF 神经网络模型, 该模型首先利用 BERT 预训练出词向量, 再将词向量输入到 BiLSTM 做进一步训练, 最后通过 CRF 解码预测最佳序列. 实验结果表明, 该模型在 MSRA 语料和人民日报语料库上分别达到了 94.65% 和 95.67% 的 $F1$ 值.

本文的创新点主要有以下两点: ① 将语言预训练模型 BERT 应用到中文实体识别中, 语言预训练是作为中文实体识别的上游任务, 它把预训练出来的结果作为下游任务 BiLSTM-CRF 的输入, 这就意味着下游主要任务是对预训练出来的词向量进行分类即可, 它不仅减少了下游任务的工作量, 而且能够得到更好的效果; ② BERT 语言预训练模型不同于传统的预训练模型, BERT 预训练出来的是动态词向量, 能够在不同语境中表达不同的语义, 相较于传统的语言预训练模型训练出来的静态词向量 (无法表征一词多义), 在中文实体识别中具有更大的优势.

2 BERT-BiLSTM-CRF 模型

2.1 模型概述

近几年来, 对于实体识别的上游任务语言预处理而言, 它一直是研究的热点问题. 而 BERT 作为先进的语言预处理模型, 可以获取高质量的词向量, 从而更有利于实体识别的下游任务进行实体提取和分类. 本文提出的 BERT-BiLSTM-CRF 模型整体结构如图 1 所示, 这个模型主要分 3 个模块. 首先标注语料经过 BERT 预训练语言模型获得相应的词向量, 之后再把词向量输入到 BiLSTM 模块中做进一步处理, 最终利用 CRF 模块对 BiLSTM 模块的输出结果进行解码, 得到一个预测标注序列, 然后对序列中的各个实体进行提取分类, 从而完成中文实体识别的整个流程.

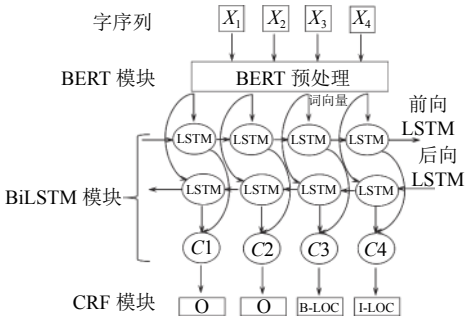

图 1 BERT-BiLSTM-CRF 模型框架

本文模型最大的优势在于 BERT 语言预处理模型的应用, 它不再需要提前训练好字向量和词向量, 只需要将序列直接输入到 BERT 中, 它就会自动提取出序列中丰富的词级特征、语法结构特征和语义特征. Ganesh Jawahar 等人[24]对 BERT 模型的内在机理做了进一步的研究, 指出对于 BERT 模型每一层学习到的特征是不尽相同的. BERT 模型的底层主要是获取短语级别的特征信息, 中层主要是学习到句法结构特征信息, 顶层则是捕获整个句子的语义信息, 经过 BERT 处理过后能够获得语境化的词向量, 对处理长距离依赖信息的语句有很好的效果. 而对于传统模型, 它们主要集中在词语或字符级别特征信息的获取, 而对于句法结构以及语义信息很少涉及. 可以看出 BERT 模型特征抽取能力明显强于传统模型.

2.2 BERT 模块

多年来, 对语言模型的研究先后经历了 one-hot、Word2Vec、ELMO、GPT 到 BERT, 前几个语言模型均存在一些缺陷, 如 Word2Vec 模型训练出来的词向量是属于静态 Word Embedding, 无法表示一词多义; GPT 则是单向语言模型, 无法获取一个字词的上下文. 而对 BERT 模型而言, 它是综合 ELMO 和 GPT 这两者的优势而构造出来的模型. Fábio Souza[22]利用 BERT 提取更强的句子语义特征来进行命名实体识别, 并取得相当不错的效果. 由于 BERT 具有很强的语义表征优势, 本文就利用 BERT 获取语境化的词向量来提高实体识别的性能. 但是本文采取的 BERT 模块与 Fábio Souza[22]有不同之处: 在对句子进行前期处理时, 他采用的是以字符为单位进行切分句子. 因此, 这样的分词方式会把一个完整的词切分成若干个子词, 在生成训练样本时, 这些被分开的子词会随机被 Mask. 而本文则按照中文的分词习惯, 于是将全词 Mask[25]的方法应

用到中文上, 在全词 Mask 中, 如果一个完整的词的部分被 Mask, 则同属该词的其他部分也会被 Mask. 具体如表 1 所示.

表 1 全词 Mask

原始文本	安徽的省会是合肥
分词文本	安徽 的 省会 是 合肥
原始 Mask 输入	[Mask]徽 的 省会 是 合[Mask]
全词 Mask 输入	[Mask][Mask] 的 省会 是 [Mask][Mask]

具体 BERT 模型结构如图 2 所示.

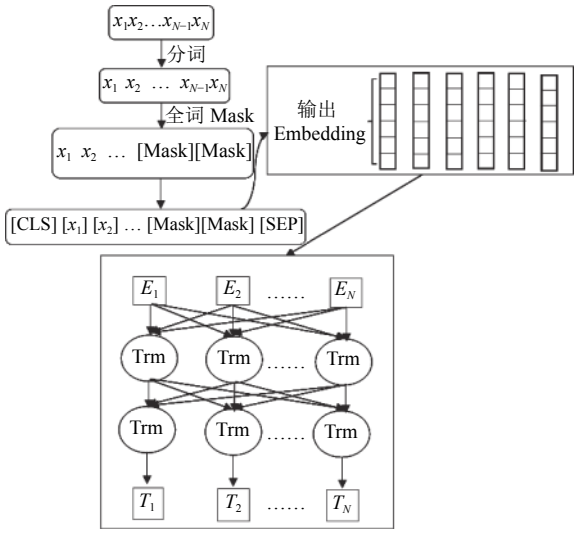

图 2 BERT 模型结构

对于任意序列, 首先通过分词处理得到分词文本序列; 然后对分词序列的部分词进行全词 Mask, 再为序列的开头添加一个特殊标记[CLS], 句子间用标记[SEP]分隔. 此时序列的每个词的输出 Embedding 由 3 部分组成: Token Embedding、Segment Embedding 和 Position Embedding. 将序列向量输入到双向 Transformer 进行特征提取, 最后得到含有丰富语义特征的序列向量.

对于 BERT 而言, 其关键部分是 Transformer 结构. Transformer 是个基于“自我注意力机制”的深度网络, 其编码器结构图如图 3 所示.

该编码器的关键部分就是自注意力机制, 它主要是通过同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:

$$Attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \quad (1)$$

其中, Q,K,V 是字向量矩阵, d_k 是 Embedding 维度. 而多头注意力机制则是通过多个不同的线性变换对 Q,K,V 进行投影, 最后将不同的 *Attention* 结果拼接起来, 公式如式 (2) 和式 (3):

$$MultiHead(Q,K,V)=Concat(head_1,\cdots,head_n)W^O \quad (2)$$

$$head_i=Attention(QW_i^Q,KW_i^K,VW_i^V) \quad (3)$$

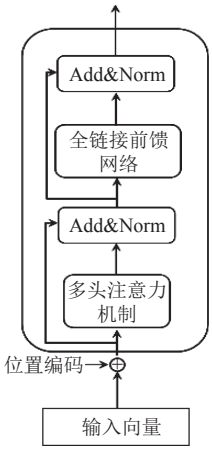

图 3 Transformer 编码器

因此模型就可以得到不同空间下的位置信息, 其中W是权重矩阵.

由于 Transformer 并没有像 RNN 一样能够获取整个句子的序列能力, 因此为解决这个问题, Transformer 在数据预处理前加入了位置编码, 并与输入向量数据进行求和, 得到句子中每个字的相对位置.

而 Transformer 结构中的全链接前馈网络有两层 dense: 第一层的激活函数是 ReLU, 第二层是一个线性激活函数. 如果多头注意力机制的输出表示为 Z, b是偏置向量, 则 *FFN* (全链接前馈网络) 可以表示为:

$$FFN(Z)=\max(0,ZW_1+b_1)W_2+b_2 \quad (4)$$

2.3 BiLSTM 模块

LSTM (Long-Short Term Memory, 长短期记忆网络), 是循环神经网络 (RNN) 的一种变体. 它解决了 RNN 训练时所产生的梯度爆炸或梯度消失. LSTM 巧妙地运用门控概念实现长期记忆, 同时它也能够捕捉序列信息. LSTM 单元结构如图 4.

LSTM 的核心主要是以下结构: 遗忘门、输入门、输出门以及记忆 Cell. 输入门与遗忘门两者的共同作用就是舍弃无用的信息, 把有用的信息传入到下一时刻. 对于整个结构的输出, 主要是记忆 Cell 的输出

和输出门的输出相乘所得到的. 其结构用公式表达如下:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\ z_t &= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\ c_t &= f_t c_{t-1} + i_t z_t \\ o_t &= \tanh(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\ h_t &= o_t \tanh(c_t) \end{aligned} \tag{5}$$

其中, σ是激活函数, W是权重矩阵, b是偏置向量, z_t是待增加的内容, c_t是t时刻的更新状态, i_t, f_t, o_t分别是输入门、遗忘门及输出门的输出结果, h_t则是整个LSTM单元t时刻的输出.

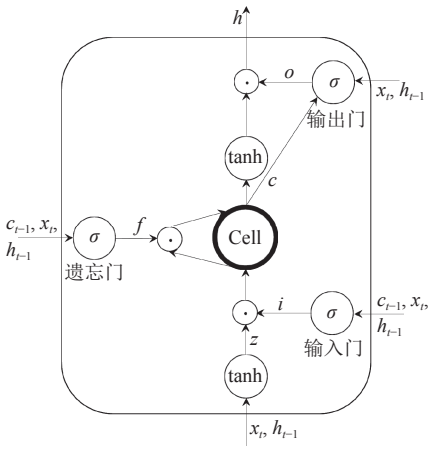

图 4 LSTM 单元结构

由于单向的 LSTM 模型无法同时处理上下文信息, 而 Graves A 等人[26]提出的 BiLSTM (Bidirectional Long-Short Term Memory, 双向长短期记忆网络), 其基本思想就是对每个词序列分别采取前向和后向 LSTM, 然后将同一个时刻的输出进行合并. 因此对于每一个时刻而言, 都对应着前向与后向的信息. 具体结构如图 5 所示. 其中输出如以下式所示:

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{6}$$

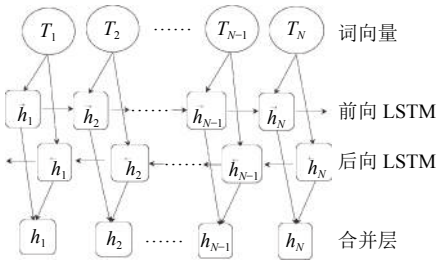

图 5 BiLSTM 模型结构

2.4 CRF 模块

在命名实体识别任务中, BiLSTM 善于处理长距离的文本信息, 但无法处理相邻标签之间的依赖关系. 而 CRF 能通过邻近标签的关系获得一个最优的预测序列, 可以弥补 BiLSTM 的缺点. 对于任一个序列 $X = (x_1, x_2, \cdots, x_n)$, 在此假定$P$是 BiLSTM 的输出得分矩阵, P的大小为$n \times k$, 其中n为词的个数, k为标签个数, P_{ij}表示第i个词的第j个标签的分数. 对预测序列 $Y = (y_1, y_2, \cdots, y_n)$而言, 得到它的分数函数为:

$$s(X, Y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{7}$$

A 表示转移分数矩阵, A_{ij}代表标签i转移为标签j的分数, A 的大小为$k+2$. 预测序列 Y 产生的概率为:

$$p(Y|X) = \frac{e^{s(X,Y)}}{\sum\limits_{\tilde{Y} \in Y_X} s(X, \tilde{Y})} \tag{8}$$

两头取对数得到预测序列的似然函数:

$$\ln(p(Y|X)) = s(X, Y) - \ln(\sum_{\tilde{Y} \in Y_X} s(X, \tilde{Y})) \tag{9}$$

式中, $\tilde{Y}$表示真实的标注序列, Y_X表示所有可能的标注序列. 解码后得到最大分数的输出序列:

$$Y^* = \arg\max_{\tilde{Y} \in Y_X} s(X, \tilde{Y}) \tag{10}$$

3 实验结果与分析

3.1 实验数据集

本文主要采用人民日报语料库和 MSRA 语料作为实验的数据集, 这两个数据集是国内公开的中文评测数据集. 它们包含了 3 种实体类型, 分别是人名、地名和组织机构. 本实验主要对人名、地名以及组织机构进行识别评测. 语料具体规模如表 2 所示.

表 2 语料规模介绍 (单位: 句)

语料	训练集	开发集	测试集
人民日报	17 573	911	1718
MSRA	46 364	---	4365

3.2 数据集标注与评价指标

命名实体识别常用的标注体系有 BIO 体系、BIOE 体系以及 BIOES 体系, 本文选用的是 BIO 体系, 该体系的标签有 7 个, 分别是“O”、“B-PER”、“I-PER”、

“B-ORG”、“I-ORG”、“B-LOC”、“I-LOC”.

本文采用召回率 R、精确率 P 和 $F1$ 值来评判模型的性能, 各评价指标的计算方法如下:

$$P=\frac{a}{B}\times 100\%$$
$$R=\frac{a}{A}\times 100\% \quad (11)$$
$$F1=\frac{2PR}{P+R}\times 100\%$$

式中, a是识别正确的实体数, A 是总实体个数, B 是识别出的实体数.

3.3 实验环境与实验参数配置

3.3.1 实验环境配置

本实验是基于 Tensorflow 平台搭建, 具体训练环境配置如表 3 所示.

表 3 训练环境配置

操作系统	Linux
CPU	Inter Xeon E5-2698 v4 2.2GHz(20-Core)
GPU	4*NVIDA Tesla V100
Python	3.5
Tensorflow	1.14.0

3.3.2 实验参数配置

训练过程中, 采用 Adam 优化器, 学习速率选取 0.001. 同时, 还设置 LSTM_dim 为 200, batch_size 为 64, max_seq_len 为 128. 为防止过拟合问题, 在 BiLSTM 的输入输出中使用 Dropout, 取值为 0.5. 具体超参数设定如表 4 所示.

表 4 参数设置

参数	值
Transformer 层数	12
隐藏层维度	768
优化器	Adam
学习速率	0.001
LSTM_dim	200
batch_size	64
max_seq_len	128
Dropout	0.5
clip	5.0

3.4 实验结果

为了对本文模型做出更加客观的评价, 本文分别对人民日报语料和 MSRA 语料进行测评, 具体实验结果如表 5 至表 8 所示 (注: 表中的 BERT-BiLSTM-CRF 指的是全词 Mask 下的 BERT-BiLSTM-CRF).

表 5 人民日报语料测试结果 (单位: %)

模型	P	R	$F1$
LSTM-CRF	84.20	80.20	82.00
BiLSTM	81.08	79.21	80.05
BiLSTM-CRF	87.21	83.21	85.09
BERT-BiLSTM-CRF(原始 Mask)	95.08	94.40	94.74
BERT-BiLSTM-CRF	**96.04**	**95.30**	**95.67**

表 6 MSRA 语料测试结果 (单位: %)

模型	P	R	$F1$
LSTM-CRF	83.45	80.84	81.48
BiLSTM	78.72	77.07	77.05
BiLSTM-CRF	86.79	84.51	85.04
BERT-BiLSTM-CRF(原始 Mask)	94.35	94.07	94.21
BERT-BiLSTM-CRF	**94.38**	**94.92**	**94.65**

表 7 训练时间 (单位: s)

模型	训练时间
LSTM-CRF	368
BiLSTM	227
BiLSTM-CRF	396
BERT-BiLSTM-CRF(原始 Mask)	1834
BERT-BiLSTM-CRF	120

表 8 在 MSRA 语料测试的模型对比 (单位: %)

模型	P	R	$F1$
DC-BiLSTM-CRF[17]	92.14	90.96	91.54
Radical-BiLSTM-CRF[16]	91.28	90.62	90.95
Lattice-LSTM-CRF[18]	93.57	92.79	93.18
WC-LSTM+pertain[19]	…	…	93.74
片段神经网络结构[20]	92.09	88.85	90.44
CNN-BiLSTM-CRF[27]	91.63	90.56	91.09
BERT-IDCNN-CRF[28]	94.86	93.97	94.41
BERT-BiLSTM-CRF	**94.38**	**94.92**	**94.65**

3.4.1 BERT-BiLSTM-CRF 和传统经典神经网络模型的对比实验

首先, 比较 LSTM-CRF 和 BiLSTM-CRF 这两者实验结果, 后者的 $F1$ 值在人民日报语料和 MSRA 语料上比前者分别高出 3.09%、3.56%. 从此可看出, BiLSTM 能够利用双向结构获取上下文序列信息, 因此效果要优于 LSTM. 其次, 比较 BiLSTM 与 BiLSTM-CRF 的实验结果, 增加 CRF 模块后, $F1$ 值在两者语料上分别提高了 5.04%、7.99%, 这主要归因于 CRF 模块能够充分利用彼此相邻标签的关联性, 像“B-PER I-ORG …”这样的标签序列无法有效地输出, 从而可以获得全局最优的标签序列, 进而能够改善实体识别性能. 随后在

BiLSTM-CRF 的基础上, 引入 BERT 模型 (原始 Mask) 进行词向量预处理, 从实验的各项指标来看, 效果相当理想, *F*1 值高达 94.74%、94.21%, 同比 BiLSTM-CRF 模型, *F*1 值已经提高了 9.65%、9.17%. 加入的 BERT 模型, 该模型可以充分提取字符级、词级、句子级甚至句间关系的特征, 从而使预训练出来的词向量能够更好地表征不同语境中的句法与语义信息, 进而增强模型泛化能力, 提高实体识别的性能. 当全词 Mask 取代原始 Mask 的 BERT 时, 在人民日报语料、MSRA 语料上分别提高了 0.93%、0.44%, 说明其提取的特征能力更强.

此外, 本文还对比分析了前 20 轮的 *F*1 值更新情况 (以人民日报测试结果为例), 如图 6 所示. 在训练初期, 两种 BERT-BiLSTM-CRF 模型就能够达到一个较高的水平, 并且会持续提升, 最后保持在相当高的水平上; 而对于传统经典神经网络模型, 在初期就处于一个相当低的水平, 只有经过多次迭代更新才会上升到一个较高的水平, 但还是无法超过 BERT-BiLSTM-CRF 模型.

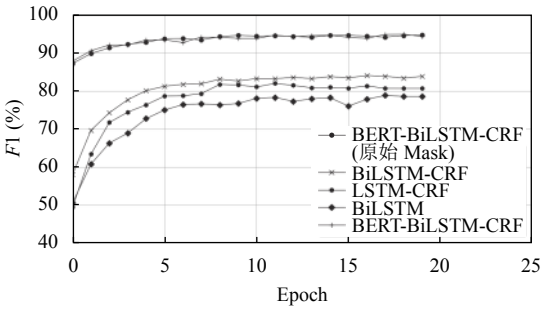

图 6 *F*1 值更新情况

同时也对比各模型训练一轮所需的时间 (以人民日报测试结果为例), 如表 7 所示.

值得比较的是后两个模型, BERT-BiLSTM-CRF (原始 Mask) 训练一轮的所需时间是本文模型的 15 倍左右, 而且本文模型的训练时间在所有模型中是最少的, 说明全词 Mask 的 BERT 具有更高的训练效率.

3.4.2 BERT-BiLSTM-CRF 和现有其他工作的对比

从表 8 中可以看出, DC-BiLSTM-CRF 模型利用 DC-BiLSTM 来学习句子特征, 应用自注意力机制来捕捉两个标注词语的关系; Radical-BiLSTM-CRF 模型使用双向 LSTM 提取字根序列的特征, 然后与字向量拼接组成模型的输入; Lattice-LSTM 模型则是把传统的 LSTM 单元改进为网格 LSTM, 然后显式地利用词与词序信息, 避免了分词错误的传递; 对于 WC-LSTM 而言, 则是利用词语信息加强语义信息, 减少分词错误的影响; 片段神经网络结构通过片段信息对片段整体进行分配标记, 从而完成实体识别. 这几种改进模型很大程度上提高了 *F*1 值.

但是上述的改进模型始终停留在对字符和词语特征的提取, 导致这些改进模型有一定的局限性. 例如, “南京市长江大桥”, 这个短语可以理解为“南京市-长江大桥”, 以也可以理解为“南京市长-江大桥”, 然而上述的模型只能获取其中的一种意思, 无法同时表征两种意思. 而本文提出的 BERT-BiLSTM-CRF 模型能很好地解决这个问题. BERT 是构建于 Transformer 之上的预训练语言模型, 它的特点之一就是所有层都联合上下文语境进行预训练. 因此 BERT 模型网络不仅可以学习到短语级别的信息表征以及丰富的语言学特征, 而且也能够学习到丰富的语义信息特征. 对于上面的“南京市长江大桥”这个例子, BERT 根据上下文不同的语境信息能够准确区分出这两种意思. 所以本文提出的 BERT-BiLSTM-CRF 与 BERT-IDCNN-CRF 模型两者相差不大, 而本文模型的 *F*1 值在 MSRA 语料上达到了 94.65%. 通过对上述多种模型的对比分析, BERT-BiLSTM-CRF 模型在所有模型中都表现出最佳的效果, 说明 BERT 相比其他模型, 其特征抽取能力更强.

4 结语

针对中文实体识别任务, 本文通过 BERT 语言预处理模型获得语境化的词向量, 再结合经典神经网络模型 BiLSTM-CRF, 构建 BERT-BiLSTM-CRF 模型. 在人民日报语料库和 MSRA 语料上分别进行评测, 相比其他模型, 本文的 BERT-BiLSTM-CRF 模型在这两者语料上都取得了最佳的结果. 本文模型, 其最大的优势在于 BERT 能够结合上下文的语义信息进行预训练, 能够学习到词级别、句法结构的特征和上下文的语义信息特征, 使得该模型相比其他模型, 具有更优的性能. 同时利用 BiLSTM 对词向量做进一步处理, 再结合 CRF 的优势, 进一步提高了中文实体识别的效果. 下一步工作可以考虑将其应用到其他领域, 进行相应的领域实体识别.

参考文献

1 彭春艳, 张晖, 包玲玉, 等. 基于条件随机域的生物命名实体识别. 计算机工程, 2009, 35(22): 197–199. [doi: 10.3969/j.issn.1000-3428.2009.22.067]

2 鞠久朋, 张伟伟, 宁建军, 等. CRF 与规则相结合的地理空间命名实体识别. 计算机工程, 2011, 37(7): 210–212, 215. [doi: 10.3969/j.issn.1000-3428.2011.07.071]

3 乐娟, 赵玺. 基于 HMM 的京剧机构命名实体识别算法. 计算机工程, 2013, 39(6): 266–271. [doi: 10.3969/j.issn.1000-3428.2013.06.059]

4 Hammerton J. Named entity recognition with long short-term memory. Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL. Stroudsburg, PA, USA. 2003. 172–175.

5 Lample G, Ballesteros M, Subramanian S, *et al*. 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, CA, USA. 2016. 260–270.

6 Pinheiro PHO, Collobert R. Recurrent convolutional neural networks for scene parsing. Proceedings of ICML. Beijing, China. 2014. 82–90.

7 Huang ZH, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv: 1508.01991, 2015.

8 Chiu JPC, Nichols E. Named entity recognition with bidirectional LSTM-CNNs. Transactions of the Association for Computational Linguistics, 2016, 4: 357–370. [doi: 10.1162/tacl_a_00104]

9 李丽双, 郭元凯. 基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别. 中文信息学报, 2018, 32(1): 116–122. [doi: 10.3969/j.issn.1003-0077.2018.01.015]

10 Luo L, Yang ZH, Yang P, *et al*. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics, 2018, 34(8): 1381–1388. [doi: 10.1093/bioinformatics/btx761]

11 Wu FZ, Liu JX, Wu CH, *et al*. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation. The World Wide Web Conference. New York, NY, USA. 2019. 3342–3348.

12 秦娅, 申国伟, 赵文波, 等. 基于深度神经网络的网络安全实体识别方法. 南京大学学报 (自然科学), 2019, 55(1): 29–40.

13 武惠, 吕立, 于碧辉. 基于迁移学习和 BiLSTM-CRF 的中文命名实体识别. 小型微型计算机系统, 2019, 40(6): 1142–1147.

14 王红斌, 沈强, 线岩团. 融合迁移学习的中文命名实体识别. 小型微型计算机系统, 2017, 38(2): 346–351.

15 王银瑞, 彭敦陆, 陈章, 等. Trans-NER: 一种迁移学习支持下的中文命名实体识别模型. 小型微型计算机系统, 2019, 40(8): 1622–1626. [doi: 10.3969/j.issn.1000-1220.2019.08.008]

16 Dong CH, Zhang JJ, Zong CQ, *et al*. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. In: Lin CY, Xue N, Zhao D, *et al*., eds. Natural Language Understanding and Intelligent Applications. Cham: Springer, 2016.239–250.

17 刘晓俊, 辜丽川, 史先章. 基于 Bi-LSTM 和注意力机制的命名实体识别. 洛阳理工学院学报 (自然科学版), 2019, 29(1): 65–70, 77.

18 Zhang Y, Yang J. Chinese NER using lattice LSTM. arXiv preprint arXiv: 1805.02023, 2018.

19 Liu W, Xu TG, Xu QH, *et al*. An encoding strategy based word-character LSTM for Chinese NER. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, MN, USA. 2019. 2379–2389.

20 王蕾, 谢云, 周俊生, 等. 基于神经网络的片段级中文命名实体识别. 中文信息学报, 2018, 32(3): 84–90, 100. [doi: 10.3969/j.issn.1003-0077.2018.03.012]

21 Devlin J, Chang MW, Lee K, *et al*. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805, 2018.

22 Souza F, Nogueira R, Lotufo R. Portuguese named entity recognition using BERT-CRF. arXiv preprint arXiv: 1909.10649, 2019.

23 Straková J, Straka M, Hajič J. Neural architectures for nested NER through linearization. arXiv preprint arXiv: 1908.06926, 2019.

24 Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy. 2019. 3651–3657.

25 Cui YM, Che WX, Liu T, *et al*. Pre-training with whole word masking for Chinese BERT. arXiv preprint arXiv: 1906.08101, 2019.

26 Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Network, 2005, 18(5–6): 602–610.

27 Jia YZ, Xu XB. Chinese named entity recognition based on CNN-BiLSTM-CRF. 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS). Beijing, China. 2018. 1–4.

28 李妮, 关焕梅, 杨飘, 等. 基于 BERT-IDCNN-CRF 的中文命名实体识别方法. 山东大学学报 (理学版), 2020, 55(1): 102–109.