doi: 10.12068/j.issn.1005-3026.2020.10.003

基于BERT模型的司法文书实体识别方法

陈 剑,何 涛,闻英友,马林涛
(东北大学 计算机科学与工程学院/东软研究院,辽宁 沈阳 110169)

摘 要: 采用手工分析案件卷宗,容易产生案件实体遗漏现象及提取特征效率低下问题.为此,使用基于双向训练Transformer的编码器表征预训练模型.在手工标注的语料库中微调模型参数,再由长短时记忆网络与条件随机场对前一层输出的语义编码进行解码,完成实体抽取.该预训练模型具有巨大的参数量、强大的特征提取能力和实体的多维语义表征等优势,可有效提升实体抽取效果.实验结果表明,本文提出的模型能实现89%以上的实体提取准确度,显著优于传统的循环神经网络和卷积神经网络模型.

关 键 词: 深度学习;预训练模型;双向长短时记忆网络;条件随机场;命名实体识别
中图分类号: TP 391 **文献标志码:** A **文章编号:** 1005-3026(2020)10-1382-06

Entity Recognition Method for Judicial Documents Based on BERT Model

CHEN Jian, HE Tao, WEN Ying-you, MA Lin-tao
(School of Computer Science & Engineering/Neusoft Research Institute, Northeastern University, Shenyang 110169, China. Corresponding author: HE Tao, E-mail: het@neusoft.com)

Abstract: Using manual analysis of case files, it is easy to cause the problem of case entity omission and low efficiency of feature extraction. Therefore, the bidirectional encoder representation from transformers pre-training model based on the traditional long short-term memory networks and conditional random fields was used to fine tune the model parameters on the manually labeled corpus for entity recognition. And then the semantic coding output from the previous layer was decoded by the long short-term memory networks and conditional random fields to complete entity extraction. The pre-training model has the advantages of huge parameters, powerful feature extraction ability and multi-dimensional semantic representation of entities, which can effectively improve the effect of entity extraction. The experimental results showed that the proposed model can achieve more than 89% entity extraction accuracy, which is significantly better than the traditional recurrent neural network and convolutional neural network model.

Key words: deep learning; pre-training model; bidirectional long short-term memory; conditional random field; named entity recognition

命名实体识别(named entity recognition, NER)问题是中文自然语言处理研究的一个重要领域,对于实体信息抽取[1]、关系抽取[2]、句法分析[3]、文本翻译[4]、知识图谱[5]构建等很多应用起到基础性作用.由于中文语言本身结构的特殊性,字词之间没有分隔符,实体描述方式多种多样,加大了中文命名实体识别的难度.

命名实体识别技术早期使用基于词典和规则的方法,根据词语的分布和语义规则进行计算打分,但需要人工完成复杂的特征建模.随着语料资源的丰富,基于统计学的机器学习算法广泛用于命名实体识别问题,典型应用包括隐马尔科夫模型[6]、最大熵模型[7]、条件随机场[8]、支持向量机[9]等.近年来,随着计算能力的提高及深度学

收稿日期: 2020-02-28
基金项目: 国家重点研发计划项目(2018YFC0830601);辽宁省重点研发计划项目(2019JH2/10100027);中央高校基本科研业务费专项资金资助项目(N171802001);辽宁省"兴辽英才计划"项目(XLYC1802100).
作者简介: 陈 剑(1982-),男,辽宁沈阳人,东北大学副教授;何 涛(1981-),男,辽宁沈阳人,东北大学东软研究院高级工程师;闻英友(1974-),男,辽宁沈阳人,东北大学教授,博士生导师.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

习的发展,卷积神经网络、循环神经网络等方法被广泛地应用在实体识别领域.

Feng 等[10]提出在嵌入词向量特征的基础上利用长短时记忆网络模型进行命名实体识别. Dong 等[11]提出利用双向长短时记忆网络模型与条件随机场进行命名实体识别. Ma 等[12]提出结合双向长短期记忆模型、卷积神经网络和条件随机场,解决序列标注问题. Strubell 等[13]在 2017 年提出迭代膨胀卷积神经网络(iterated dilated convolutional neural networks,IDCNN)来处理序列问题,IDCNN 计算出每个词分类的概率,而条件随机场(conditional random field,CRF)层引入序列的转移概率. Vaswani 等[14]提出基于多头自注意力机制的 Transformer 模型,提高了文本特征提取能力,为序列标注任务提出新的解决方法. Zhang 等[15]提出了一种新型的 Lattice LSTM(long short-term memory),将潜在的词语信息融合到基于字模型的传统 LSTM + CRF 中去,而其中潜在的词语信息是通过外部词典获得的. Yang 等[16]提出了一种利用众包标注数据学习对抗网络模型的方法,构建中文实体识别系统. 2018 年 10 月底,Google 公布 BERT(bidirectional encoder representation from transformers)[17]在 11 项 NLP 任务中刷新纪录,BERT 的成功引起业界的广泛关注.

司法文书命名实体识别是司法业务信息化和智能化的基础任务,是知识图谱构建、案情辅助研判、类案检索、法律法规推荐等上层功能的前提工作. 目前司法文书命名实体识别的研究并不成熟,也没有公开的司法领域命名实体标注数据集,因此,开发司法实体标注工具和根据实际需求标注数据集也是本文需要完成的工作.

司法文书不同于普通文本,通常包含大量的人名,如被告人、受害人、证人、代理人等多种类型,并常常使用代称; 专业术语较多,通常会出现法律法规条文; 还要关注多义词问题,如“盗窃车钥匙一把”和“一把夺过行人的背包”,两个“一把”含义不同,归入的实体类型也不同.

为了解决上述问题,本文引入 BERT 模型,该模型是一个强大的预训练模型,通过双向训练 Transformer 编码器从海量的无标注语料中学习短语信息特征、语言学特征和一定程度的语义信息特征. BERT 可以将丰富的语言学知识进行迁移学习,在规模较小的司法文书标注语料库上进行微调,同时其强大的词向量表征能力能够有效区分多义词在不同上下文中的含义.

本文提出一种融合 BERT 的多层次司法文书实体识别模型,实验结果表明,该模型与目前主流的实体识别模型 BiLSTM(bi-directional LSTM) + CRF 相比,F_1 分数可提高 7.5% 左右.

1　构建 BERT + BiLSTM + CRF 模型

BERT + BiLSTM + CRF 模型的整体结构见图 1. 三层结构分别是: ①BERT 使用 Transformer 机制对输入数据进行编码,使用预训练模型获取字的语义表示; ②BiLSTM 在 BERT 输出结果的基础上进一步提取数据的高层特征; ③CRF 对 BiLSTM 层的输出结果进行状态转移约束.

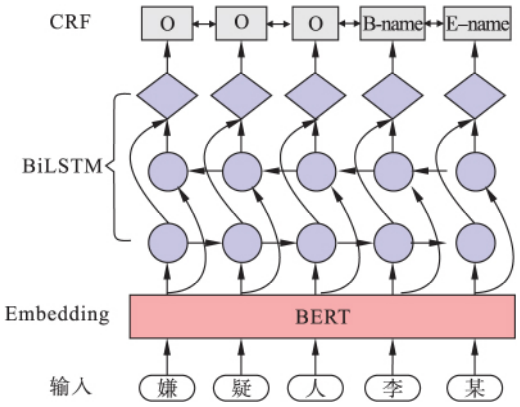

图 1　BERT + BiLSTM + CRF 模型整体架构

Fig. 1　Structure of BERT + BiLSTM + CRF model

1.1　BERT 预训练模型

BERT 预训练模型与其他词向量预训练模型如 ELMO[18],GPT 等不同,该模型是在多层 Transformer 编码器的基础上实现的. Transformer 编码器作为文本特征提取器,很多研究证明其特征提取能力远远大于 RNN 和 CNN 模型,这也是 BERT 模型的核心优势所在.

Transformer 是由 Vaswani 等[14]提出的一个完全依赖自注意力机制计算输入和输出的表示,而不使用序列对齐的递归神经网络或卷积神经网络的转换模型. 自注意力的计算方法如下: 从编码器的每个输入向量中创建三个向量: 一个 Query 向量、一个 Key 向量和一个 Value 向量. 这些向量是通过将词嵌入向量与三个训练后的矩阵 $\boldsymbol{W}_q$, $\boldsymbol{W}_k$, $\boldsymbol{W}_v$ 相乘得到的,维度默认为 64. 为了便于计算,将三个向量分别合并成矩阵,得到自注意力层的计算公式:

$$\boldsymbol{Z} = \text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}. \tag{1}$$

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

式中:Q 为 Query 向量组合的矩阵;K 为 Key 向量组合的矩阵;V 为 Value 向量组合的矩阵;d 为 Query 向量的维度,除以 $\sqrt{d_k}$ 可以使训练过程中的梯度下降更加稳定. 由 Softmax 将分数进行归一化,每个单词的得分决定了对某个位置上的单词进行编码时,对其他单词的关注程度. 由于 BERT 的目标是生成语言模型,只需要使用 Transformer 的编码器机制,因此对 Transformer 的解码器部分涉及的内容不再赘述.

在 Transformer 的基础上,BERT 使用 Masked LM 进行无监督预训练. 一个深度双向模型要比单向的"左-右"模型或浅层融合"左-右"和"右-左"模型更高效. 为了解决双向训练中每个词在多次上下文可以间接看见自己的问题,BERT 采用随机遮掩一定百分比的输入 token,然后通过预测被遮掩的 token 进行训练.

BERT 的输入数据是基于字符级 Embedding 的线性序列,每个序列的第一个 token 是一个特殊的分类识符号,记作"[CLS]",序列之间使用分隔符"[SEP]"分割. 每个字符有三个 Embedding:①Token Embedding,即每一个输入字符的 Embedding;②Segment Embedding,BERT 是一个句子级别的语言模型,这个标记对应一个句子的唯一向量表示;③Position Embedding,在自然语言处理任务中,序列的索引信息很重要. 与 Transformer 不同,BERT 并没有采用三角函数来表达句子中词语位置的方法,而是直接设置句子的固定长度去训练 Position Embedding,在每个词的位置随机初始化词向量. 最终把单词对应的三个 Embedding 叠加,形成 BERT 的输入,如图 2 所示.

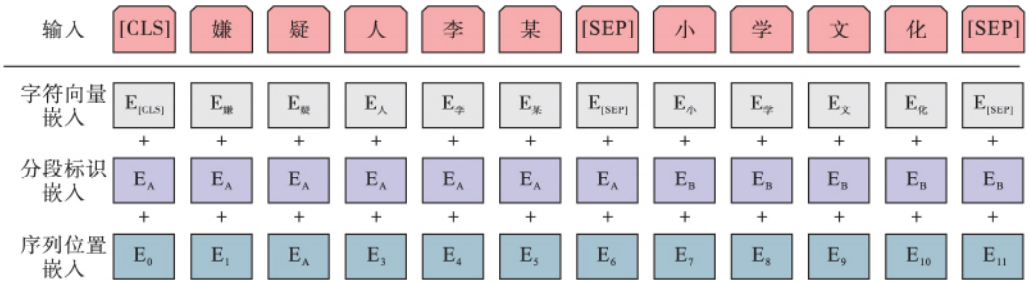

图2 BERT 模型的输入表征

Fig. 2 Input representations of BERT model

1.2 BiLSTM 层

由于在深度神经网络中使用反向传播算法,即根据损失函数计算的误差通过梯度反向传播的方式,指导深度网络权值的更新优化,随着网络深度的增加,常常出现梯度消失和梯度爆炸问题[19]. LSTM 在 RNN 的基础上进行改进,通过使用复杂的门机制解决了梯度消失/爆炸问题,加快训练收敛速度,并能很好地检测出序列中的长距离依赖. 一个基本的 LSTM 单元结构如图 3 所示. 图中:$h_{(t)}$ 存储短期状态信息,$c_{(t)}$ 存储长期状态信息. $c_{(t-1)}$ 进入时间迭代 t 的神经元时,首先经过一个忘记门限,丢掉一些记忆,再通过输入门限有选择地增加一些新记忆,$c_{(t)}$ 作为长期状态被传入下一个时间迭代 $t+1$ 的神经元. 同时,长期状态被复制并传入 tanh 函数,然后结果被输出门限过滤,产生短期状态 $h_{(t)}$:

$$i_{(t)} = \sigma(W_{xi}^{\mathrm{T}} \cdot X_{(t)} + W_{hi}^{\mathrm{T}} \cdot h_{(t-1)} + b_i), \quad (2)$$

$$f_{(t)} = \sigma(W_{xf}^{\mathrm{T}} \cdot X_{(t)} + W_{hf}^{\mathrm{T}} \cdot h_{(t-1)} + b_f), \quad (3)$$

$$o_{(t)} = \sigma(W_{xo}^{\mathrm{T}} \cdot X_{(t)} + W_{ho}^{\mathrm{T}} \cdot h_{(t-1)} + b_o), \quad (4)$$

$$g_{(t)} = \tanh(W_{xg}^{\mathrm{T}} \cdot X_{(t)} + W_{hg}^{\mathrm{T}} \cdot h_{(t-1)} + b_g), \quad (5)$$

$$c_{(t)} = f_{(t)} \otimes c_{(t-1)} + i_{(t)} \otimes g_{(t)}, \quad (6)$$

$$y_{(t)} = h_{(t)} = o_{(t)} \otimes \tanh(c_{(t)}). \quad (7)$$

式中:W_{xi},W_{xf},W_{xo},W_{xg} 为每一层连接到输入向量 $x_{(t)}$ 的权重矩阵;W_{hi},W_{hf},W_{ho},W_{hg} 为每一层连接到前一个短期状态 $h_{(t-1)}$ 的权重矩阵;b_i,b_f,b_o,b_g 是每一层的偏差系数.

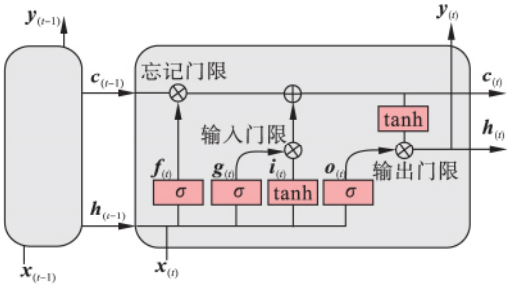

图3 LSTM 单元结构

Fig. 3 LSTM unit structure

在处理序列标注问题时,神经网络模型不仅要关注上文信息,同样也要关注下文信息,将前向 LSTM 和后向 LSTM 结合起来,使得每一个训练序列向前和向后分别是两个循环神经网络,而且这两个网络连接着同一个输出层,这便是

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

BiLSTM 的优点,能够提供给输出层输入序列中每一个元素完整的上下文信息.

1.3　CRF 层

CRF 是一种判别式概率图模型,在序列标注任务中通常使用线性链条件随机场. 设随机变量序列 x 为观测序列,y 为状态向量(标记序列),每一个(y_i,x_i)对为一个线性链上的最大团,并满足:

$$P(y_i|\boldsymbol{x},y_1,y_2,\cdots,y_n)=P(y_i|\boldsymbol{x},y_{i-1},y_{i+1}). \tag{8}$$

给定一条观测序列 $\boldsymbol{x}$,使用 CRF 求解状态序列 y 的建模公式为

$$P(\boldsymbol{y}|\boldsymbol{x})=\frac{1}{Z(\boldsymbol{x})}\prod_i\exp\Big(\sum_k\lambda_k f_k(y_{i-1},y_i,\boldsymbol{x},i)\Big)=\frac{1}{Z(\boldsymbol{x})}\exp\Big(\sum_{i,k}\lambda_k f_k(y_{i-1},y_i,\boldsymbol{x},i)\Big). \tag{9}$$

因为状态序列与前后 Token 之间存在限定关系,并与观测序列存在依赖关系,所以引入两类特征函数,转移特征函数集 t 和状态特征函数集 s,建模公式可扩展为

$$P(\boldsymbol{y}|\boldsymbol{x})=\frac{1}{Z(\boldsymbol{x})}\exp\Big(\sum_{i,j}\lambda_j t_j(y_{i-1},y_i,\boldsymbol{x},i)+\sum_{i,l}\mu_l s_l(Y_i,\boldsymbol{x},i)\Big), \tag{10}$$

$$Z(\boldsymbol{x})=\sum_y\exp\Big(\sum_{i,j}\lambda_j t_j(y_{i-1},y_i,\boldsymbol{x},i)+\sum_{i,l}\mu_l s_l(y_i,\boldsymbol{x},i)\Big). \tag{11}$$

式中: $Z(x)$ 用来归一化; t_j 表示转移状态函数,对应权重为 λ_j; s_l 为状态特征函数,对应权重为 μ_l; k 和 l 为特征函数的个数.

t_k 和 s_l 的取值为 1,0,以 t_k 为例,公式为

$$t_k(y_{i-1},y_i,\boldsymbol{x},i)=\begin{cases}1, & y_{i-1},y_i,\boldsymbol{x}\text{ 的值符合条件;}\\ 0, & \text{其他.}\end{cases} \tag{12}$$

CRF 是全局范围内统计归一化的条件状态转移概率矩阵,让底层深度神经网络在 CRF 的特征限定下,依照新的损失函数,学习出一套更合理的非线性变换空间.

2　实验结果分析

2.1　语料库收集与标注

本文所使用的标注语料库均是自主完成创建,所用语料均来自中国裁判文书网公开的裁判文书,搜集到的文书总数量达到 10 万份. 司法文书实体提取问题,不同于传统的 NER 是对人名、地名、组织机构名等 7 类特定实体的识别,案件要素提取需要自己定义复杂的实体标签集,针对不同的犯罪类型,标签集包含的内容也各不相同. 以盗窃罪为例,需要定义的实体有: 公诉机关、被告人、犯罪时间、犯罪地点、作案工具、盗窃方式、盗窃物品、被盗物品价值、盗窃金额、销赃处置、销赃金额、抓获时间、赃物追回、认罪态度、谅解情况等共 15 类实体.

标注语料库的质量对训练深度学习模型的性能起到决定性作用,本实验从语料库中选取盗窃类型且内容较为详实的 2 900 份文书进行标注,所有的标注工作均由经过专业培训的人员手工标注完成. 尽管不排除主观因素对实体标注边界的影响,但总体而言标注质量较高,非常适合用于模型的训练.

对于标注完成的语料,使用包含信息量更丰富的 BIOES[20] 标注体系转化为可供模型使用的数据集,即一个字符对应一个标签. B 表示一个实体的开始,I 表示实体内部,O 表示不是标注的实体,E 表示实体的结束,S 表示这个字本身就是一个单独的实体. 显然,标签之间存在顺序关系,比如在 B 之后不会出现 O 和 S 标签,在 I 之后只能出现 I 和 E 标签. CRF 可以约束标签之间的依赖关系.

从数据集中随机抽取三部分作为训练集、验证集、测试集,三者的文书数量比例约为 4∶1∶1. 样本分布数据如表 1 所示.

表 1　盗窃案实体样本分布

Table 1　Entity sample distribution of theft

项目	训练集	验证集	测试集
文书数量	1 900	500	500
实体数量	57 168	15 362	14 925

2.2　结果对比

在机器学习中评估模型的性能通常使用精度 P、召回率 R,F_1 分数三个指标:

$$P=\mathrm{TP}/(\mathrm{TP}+\mathrm{FP}), \tag{13}$$

$$R=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN}), \tag{14}$$

$$F_1=2\times P\times R/(P+R). \tag{15}$$

式中: TP 表示真正类的数量; FP 表示假正类的数量; FN 表示假负类的数量; F_1 分数是精度 P 和召回率 R 的谐波平均值,只有当召回率和精度都很高时,才能获得较高的 F_1 分数.

在验证集上各实体的 F_1 分数随迭代次数而发生变化的情况如表 2 所示.

设置训练过程的 epoch 最大值为 100,在每个 epoch 运行完以后,通过在验证集上综合全部实体识别的得分来获取性能最优的模型. 训练完成

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表2 单次训练过程中各实体识别 F_1 分数随迭代次数的变化情况

Table 2 Change of F_1 score of entity recognition with the number of iterations in the process of training

实体类型	迭代次数									
	5	10	15	20	25	30	35	40	45	50
公诉机关	0.705 2	0.836 5	0.812 1	0.885 5	0.861 7	0.836 1	0.897 1	0.932 1	0.953 6	0.956 0
被告人	0.492 9	0.635 4	0.675 4	0.525 7	0.745 1	0.808 3	0.863 6	0.899 6	0.920 5	0.926 3
犯罪时间	0.802 2	0.830 7	0.827 1	0.837 5	0.860 1	0.887 2	0.900 5	0.891 6	0.917 1	0.929 1
犯罪地点	0.618 8	0.677 6	0.693 1	0.709 2	0.734 6	0.809 8	0.826 7	0.877 3	0.884 6	0.892 2
作案工具	0.506 0	0.582 2	0.656 7	0.651 2	0.637 7	0.683 7	0.753 5	0.816 4	0.852 9	0.876 2
盗窃方式	0.388 9	0.419 5	0.515 6	0.495 2	0.569 2	0.626 2	0.663 1	0.712 0	0.781 1	0.802 5
盗窃物品	0.477 7	0.627 4	0.751 3	0.799 3	0.789 8	0.831 6	0.880 2	0.892 0	0.924 7	0.926 4
被盗物品价值	0.526 0	0.586 3	0.627 5	0.735 3	0.736 7	0.785 0	0.838 0	0.888 6	0.909 6	0.919 1
盗窃金额	0.741 0	0.787 3	0.820 0	0.833 5	0.832 4	0.871 7	0.868 0	0.888 5	0.908 5	0.919 2
销赃处置	0.413 0	0.595 4	0.543 6	0.652 5	0.628 5	0.699 3	0.682 1	0.756 8	0.787 1	0.796 0
销赃金额	0.362 6	0.493 6	0.529 0	0.581 7	0.631 1	0.673 2	0.765 1	0.821 5	0.862 1	0.876 8
抓获时间	0.506 0	0.582 2	0.656 7	0.651 2	0.737 7	0.783 7	0.753 5	0.816 4	0.872 9	0.896 2
赃物追回	0.292 1	0.364 0	0.439 7	0.426 5	0.558 3	0.633 2	0.692 5	0.756 2	0.793 6	0.806 0
认罪态度	0.411 3	0.640 7	0.650 0	0.696 1	0.723 4	0.734 3	0.765 2	0.799 0	0.814 1	0.833 3
谅解情况	0.264 2	0.529 4	0.470 6	0.571 4	0.694 6	0.729 4	0.794 6	0.726 3	0.765 2	0.780 0

后,使用该模型在测试集上获得各类实例的性能指标,其指标分布如表3所示.

表3 15类实体的识别效果

Table 3 Recognition effect of 15 kinds of entities

实体	P	R	F_1
公诉机关	0.953 5	0.960 4	0.956 9
被告人	0.941 5	0.930 8	0.936 1
犯罪时间	0.922 2	0.924 7	0.923 5
犯罪地点	0.886 6	0.915 6	0.900 9
作案工具	0.856 0	0.872 4	0.864 1
盗窃方式	0.808 4	0.864 7	0.835 6
盗窃物品	0.928 7	0.925 0	0.926 9
被盗物品价值	0.927 6	0.909 0	0.918 2
盗窃金额	0.902 5	0.940 7	0.921 2
销赃处置	0.801 6	0.818 8	0.810 1
销赃金额	0.889 2	0.894 5	0.891 9
抓获时间	0.924 9	0.874	0.898 7
赃物追回	0.800 3	0.832 0	0.815 8
认罪态度	0.802 4	0.876 4	0.837 8
谅解情况	0.757 7	0.877 6	0.813 2

该模型在各个实体类型上的得分会有较大的差异,比如"盗窃方式"的 F_1 分数只有0.835 6,其原因是该类型实体在概念上不够明确,盗窃方式的描述多种多样,很难确定每种盗窃方式的边界,在标注的时候不像"盗窃时间"有明确的边界,导致在标注的过程中受到较大的主观因素和标注习惯影响;而"销赃处置"的 F_1 分数较低是因为该类实体数量较少,模型学习得不够充分.

为了证明本文提出的模型在性能方面的优越性,在相同数据集上,与以下模型进行比较:

1) IDCNN + CRF 模型:迭代扩张卷积神经网络通过融合空洞卷积方法,实现序列化特征的卷积提取与上下文特征传递.

2) BiLSTM + CRF 模型:序列标注问题使用主流模型,使用 BiLSTM 获取文本特征并输出各个标签取值的概率,CRF 对标签间的顺序关系做约束.

3) BERT + CRF 模型:通过每个字符左边和右边的上下文信息,BERT 层可以学到句子中每个字符最可能对应的实体标注是什么,CRF 可以调整违反标注规则的结果,降低错误率.

本文将命名实体识别领域广泛使用的三个模型与 BERT + BiLSTM + CRF 模型的性能指标进行统计对比,各模型性能如表4所示.

表4 4种模型结果对比

Table 4 Comparison of results of 4 models

模型	P	R	F_1
IDCNN + CRF	0.817 4	0.786 4	0.801 6
BiLSTM + CRF	0.808 5	0.824 4	0.816 4
BERT + CRF	0.889 2	0.872 1	0.880 5
BERT + BiLSTM + CRF	0.880 5	0.902 2	0.891 2

从统计数据可以看出,前三个模型的 F_1 分数

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

分别达到了 0.801 6,0.816 4,0.880 5,相差并不明显. 本文提出的模型 F_1 分数达到了 0.891 2,比目前广泛应用的 BiLSTM + CRF 提高 7.5%,从另一个角度来看,实体识别的错误率下降 40%,性能有了显著提升. BERT + BiLSTM + CRF 模型比 BERT + CRF 模型的 F_1 分数高出 1.07%,两者的结果非常接近,这也说明 BERT 的特征提取能力非常强大,在计算资源有限的情况下,BERT + CRF应用于实体识别也是一种很好的选择.

3 结　论

1) 为解决检察业务标注语料规模较小、案件文本实体提取困难的问题,提出一种将 BERT,BiLSTM 与 CRF 相结合,进行案件实体提取的方法. 该方法利用 BERT 预训练学习的语义句法知识,在标注的数据集上进行微调,再利用 BiLSTM 的序列建模能力和 CRF 的状态转移约束功能进一步优化. 实验表明,该方法优于目前几种主流的实体识别模型,在性能上得到大幅提升.

2) 下一步的任务是如何将该模型应用到检察案件文本的智能标注上及如何在云平台上提供高效可信的文本标注服务和实体提取服务.

参考文献:

[1] Sundar N G, Sunny A T. An efficient information extraction model for personal named entity [J]. *International Journal of Computer Trends & Technology*, 2013, 4(3): 119 – 128.

[2] Li D, Huang L, Ji H, et al. Biomedical event extraction based on knowledge-driven tree-LSTM [C] //Proceedings of NAACL-HLT 2019. Minneapolis, 2019: 1421 – 1430.

[3] 甘丽新,万常选,刘德喜,等. 基于句法语义特征的中文实体关系抽 [J]. 计算机研究与发展,2016,53(2): 284 – 302.
(Gan Li-xin, Wan Chang-xuan, Liu De-xi, et al. Extraction of Chinese entity relations based on syntactic and semantic features [J]. *Journal of Computer Research and Development*, 2016, 53(2): 284 – 302.)

[4] Nikoulina V, Sandor A, Dymetman M. Hybrid adaptation of named entity recognition systems for statistical machine translation purposes [J]. *Journal of Radiation Research*, 2011, 53(2): 1 – 16.

[5] 王鑫,邹磊,王朝坤,等. 知识图谱数据管理研究综述 [J]. 软件学报,2019,30(7): 2139 – 2174.
(Wang Xin, Zou Lei, Wang Chao-kun, et al. A survey of knowledge map data management [J]. *Journal of Software*, 2019, 30(7): 2139 – 2174.)

[6] 俞鸿魁,张华平,刘群,等. 基于层叠隐马尔可夫模型的中文命名实体识别 [J]. 通信学报,2006,27(2): 87 – 94.
(Yu Hong-kui, Zhang Hua-ping, Liu Qun, et al. Chinese named entity recognition based on cascading hidden Markov model [J]. *Journal on Communications*, 2006, 27(2): 87 – 94.)

[7] McCallum A, Freitag D, Pereira F C N. Maximum entropy Markov models for information extraction and segmentation [C] //Proceedings of the Seventeenth International Conference on Machine Learning. Sydney, 2000: 591 – 598.

[8] 周俊生,戴新宇,尹存燕,等. 基于层叠条件随机场模型的中文机构名自动识别 [J]. 电子学报,2006,34(5): 804 – 809.
(Zhou Jun-sheng, Dai Xin-yu, Yin Cun-yan, et al. Automatic recognition of Chinese organization name based on cascading conditional random field model [J]. *Acta Electronica Sinica*, 2006, 34(5): 804 – 809.)

[9] Ertekin S, Bottou L. Nonconvex online support vector machines [J]. *Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(2): 368 – 381.

[10] Feng Y H, Hong Y U, Sun G, et al. Named entity recognition method based on BLSTM [J]. *Computer Science*, 2018, 45(2): 261 – 268.

[11] Dong C, Zhang J, Zong C, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition [C] //Natural Language Processing and Chinese Computing 2016. Kunming, 2016: 239 – 250.

[12] Ma X, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF [C] //Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, 2016: 1064 – 1074.

[13] Strubell E, Verga P, Belanger D, et al. Fast and accurate entity recognition with iterated dilated convolutions [C] //Conference on Empirical Methods in Natural Language Processing. Copenhagen, 2017: 2670 – 2680.

[14] 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. New York, 2017: 6000 – 6010.

[15] Zhang Y, Yang J. Chinese NER using lattice LSTM [C] //Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, 2018: 1554 – 1564.

[16] Yang Y S, Zhang M S, Chen W L, et al. Adversarial learning for Chinese NER from crowd annotations [C] //Proceedings of the 32th AAAI Conference on Artificial Intelligence. New Orleans, 2018: 3216 – 3222.

[17] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C] //Proceedings of the 2019 Conference of the Association for Computational Linguistics. Stroudsburg, 2019: 4171 – 4186.

[18] Matthew E P, Mark N, Mohit I, et al. Deep contextualized word representations [C] //Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics. New Orleans, 2018: 3253 – 3268.

[19] Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep network [J]. *Advances in Neural Information Processing System*, 2007, 6(19): 153 – 162.

[20] Li J Q, Zhao S H, Yang J J. WCP-RNN: a novel RNN-based approach for Bio-NER in Chinese EMRs [J]. *Journal of Supercomputing*, 2018, 1(16): 1 – 18.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net