智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

加入自注意力机制的BERT命名实体识别模型

毛明毅， 吴晨， 钟义信， 陈志成

引用本文:
毛明毅, 吴晨, 钟义信, 等. 加入自注意力机制的BERT命名实体识别模型[J]. 智能系统学报, 2020, 15(4): 772–779.
MAO Mingyi, WU Chen, ZHONG Yixin, et al. BERT named entity recognition model with self-attention mechanism[J]. *CAAI Transactions on Intelligent Systems*, 2020, 15(4): 772–779.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202003003

DOI: 10.11992/tis.202003003

加入自注意力机制的 BERT 命名实体识别模型

毛明毅[1], 吴晨[1], 钟义信[2], 陈志成[2]

(1. 北京工商大学 计算机与信息工程学院, 北京 100048; 2. 北京邮电大学 计算机学院, 北京 100876)

摘 要: 命名实体识别属于自然语言处理领域词法分析中的一部分, 是计算机正确理解自然语言的基础。为了加强模型对命名实体的识别效果, 本文使用预训练模型 BERT(bidirectional encoder representation from transformers) 作为模型的嵌入层, 并针对 BERT 微调训练对计算机性能要求较高的问题, 采用了固定参数嵌入的方式对 BERT 进行应用, 搭建了 BERT-BiLSTM-CRF 模型。并在该模型的基础上进行了两种改进实验。方法一, 继续增加自注意力 (self-attention) 层, 实验结果显示, 自注意力层的加入对模型的识别效果提升不明显。方法二, 减小 BERT 模型嵌入层数。实验结果显示, 适度减少 BERT 嵌入层数能够提升模型的命名实体识别准确性, 同时又节约了模型的整体训练时间。采用 9 层嵌入时, 在 MSRA 中文数据集上 F_1 值提升至 94.79%, 在 Weibo 中文数据集上 F_1 值达到了 68.82%。

关键词: 命名实体识别; BERT; 自注意力机制; 深度学习; 条件随机场; 自然语言处理; 双向长短期记忆网络; 序列标注

中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2020)04−0772−08

中文引用格式: 毛明毅, 吴晨, 钟义信, 等. 加入自注意力机制的 BERT 命名实体识别模型 [J]. 智能系统学报, 2020, 15(4): 772–779.

英文引用格式: MAO Mingyi, WU Chen, ZHONG Yixin, et al. BERT named entity recognition model with self-attention mechanism[J]. CAAI transactions on intelligent systems, 2020, 15(4): 772–779.

BERT named entity recognition model with self-attention mechanism

MAO Mingyi[1], WU Chen[1], ZHONG Yixin[2], CHEN Zhicheng[2]

(1. School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China; 2. School of Computer, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Named entity recognition is a part of lexical analysis in the field of natural language processing. It is the basis for a computer to correctly understand natural language. In order to strengthen the recognition effect of the model on named entities, in this study, the pre-trained model BERT (bidirectional encoder representation from transformers) was used as the embedding layer of the model, and fixed parameter embedding was adopted to solve the problem of high computer performance required for BERT fine-tuning training. A BERT-BiLSTM-CRF model was built, and on the basis of this model, two improved experiments were carried out. Method one is to continue to add a self-attention layer. Experimental results show that the addition of the self-attention layer does not significantly improve the recognition effect of the model. Method two is to reduce the number of embedding layers of the BERT model. Experimental results show that moderately reducing the number of BERT embedding layers can improve the model's named entity recognition accuracy, while saving the overall training time of the model. When using 9-layer embedding, thevalue on the MSRA Chinese data set increased to 94.79%, and thevalue on the Weibo Chinese data set reached 68.82%.

Keywords: named entity recognition; bidirectional encoder representation from transformers; self-attention mechanism; deep learning; conditional random field; natural language processing; bi-directional long short-term memory; sequence tagging

收稿日期: 2020−03−02.
基金项目: 北京市自然科学基金项目 (4202016).
通信作者: 毛明毅. E-mail: maomy@th.btbu.edu.cn.

命名实体识别 NER(named entity recognition) 是自然语言处理研究领域的基础性工作之一, 任

务是提取非结构化文本中具有特定含义的实体。命名实体能否准确识别对于自然语言处理上层任务包括信息检索、自动问答、信息抽取、知识图谱构建都有着重要的影响。对于分词、词性标注等底层序列标注任务, 也存在着相互影响, 命名实体识别错误将导致分词错误, 进而导致词性标注错误, 反之如果利用基于词的命名实体识别方法, 分词错误又将导致命名实体识别错误。在现在这个数据量飞速增长的数字时代, 从海量数据中快速筛选有用信息, 准确获取命名实体是其中关键性的一步。

1　相关工作

对命名实体识别的研究最早可以追溯到 1991 年在 IEEE 人工智能会议上发表的一篇关于识别公司名称的文章[1]。1996 年第 6 届消息理解会议 (Message Understanding Conference, MUC6) 上正式将命名实体识别学术定义, 并列为研究课题。最初的命名实体识别实现多基于规则模板, 语言学家通过语言学知识, 构造识别规则匹配模板。这样的识别方法不仅耗费大量专业人员人力, 而且可移植性较差。更换领域甚至更换语境都会对识别准确率造成较大影响。

文献 [2] 首先在 1999 年提出了基于隐马尔可夫模型的命名实体识别方法, 开启了基于概率统计方法进行命名实体识别的新时期。此后, 最大熵模型、支持向量机[3]、条件随机场模型[4] 纷纷在命名实体识别领域有所应用, 其中条件随机场模型识别效果最佳, 直到现在依然是主流命名实体识别模型的重要组成部分。

近年来, 随着计算机性能的快速提升, 深度学习逐渐成为研究热点。基于深度学习的实体识别方法也渐渐成为主流研究方法。文献 [5] 提出采用卷积神经网络实现命名实体识别。文献 [6] 提出利用双向长短期记忆网络实现命名实体识别, 并在 CoNLL2003 数据集上取得了 84.74% 的当时最好成绩。文献 [7] 通过在模型中添加用于提取单词字符级表示的卷积神经网络, F_1 值达到了 91.21%。文献 [8] 通过在 BiLSTM-CRF 模型中加入注意力机制, 对命名实体识别效果做了进一步改善。

在中文命名实体识别领域, 近期研究主要分为 3 个方向:

1) 通过迁移学习、对抗学习、远程监督等方式降低命名实体识别训练语料标注工作量。文献 [9] 利用了远程监督的方法。文献 [10] 希望通过主动学习在少量标注的情况下使命名实体识别达到较好的效果。文献 [11] 利用 K 近邻解决标注语料少的问题。文献 [12] 通过 AdaBoost 算法实现迁移学习, 同样希望降低人工标注的工作量。

2) 专业领域命名实体识别应用, 将常规命名实体识别模型应用到特定领域中。文献 [13] 研究了面向国防科技领域术语的识别技术。文献 [14] 研究了面向电力问答领域命名实体识别技术。文献 [15] 研究了植物命名实体识别技术。

3) 改进命名实体识别模型, 提高模型学习效果。文献 [16] 提出基于 Lattice LSTM 的命名实体识别方法, 在 MSRA 数据集上成绩为 93.18%, 但是由于模型中词嵌入长度不同, 导致无法并行训练, 模型训练缓慢。文献 [17] 提出的引入自注意力机制的中文命名实体识别方法, 在 Weibo 数据集上的成绩为 58.76%。文献 [18] 利用与中文分词对抗学习改进命名实体识别准确度。

现有模型存在的主要问题有两个方面, 一方面是传统模型 (如 BiLSTM-CRF) 在识别准确性上还有较大提升空间, 另一方面在准确性较高的模型上普遍存在着训练效率低或训练成本高的问题, 很难在实际项目中得到应用, 如 Lattice LSTM 模型、BERT 微调模型。

因此本文在第 3 个研究方向上做了进一步研究, 首先根据文献 [19] 的研究结果显示, BERT 在命名实体识别任务应用上, 不做微调与微调两种应用方式的结果准确性差距与在其他应用领域相比较小。而微调的训练成本较高, 所以本文首先采用了固定 BERT 参数 BERT-BiLSTM-CRF 的 3 层模型。之后尝试了 2 种方式对模型做进一步改进。由于固定了 BERT 参数不进行参数微调, 所以本文首先尝试了在 3 层模型的基础上继续增加自注意力 (self-attention) 层是否能够提升模型实体识别准确性, 之后又尝试了减少 BERT 层数是否会对模型识别准确率造成影响。

2　本文模型框架

本文基础模型框架为固定 BERT 参数 BERT-BiLSTM-CRF 的 3 层模型, 由 BERT 嵌入层、BiLSTM 双向长短期记忆网络层和 CRF 条件随机场层所组成, 模型整体结构如图 1 所示。

2.1　BERT 嵌入层

2018 年, 谷歌人工智能团队提出了 BERT 预训练语言模型[20], 在 11 个自然语言处理任务中刷新了当时最好成绩。运用双向 Transformer 神经

网络作为编码器，使得对每个字的预测可以参考前后双方向的文字信息。模仿中国英语考试中的完型填空题，随机掩盖部分输入词，通过句子中其他词语对被掩盖词进行预测。除此之外模型训练中还增加了一个句子级别上连续性预测的创新任务。并且在谷歌为机器学习定制的专用芯片 TPU 上进行了海量数据预训练，后续预训练模型多以 BERT 模型为基础进行改进，包括 XLNet[21]、RoBERTa、ALBERT 等，这些模型在自然语言处理领域的其他研究中，如阅读理解、问题匹配、语言推断等问题上的解决效果均有不同程度提升，但是通过实验发现在序列标注问题上，效果均不如 BERT 预训练模型。

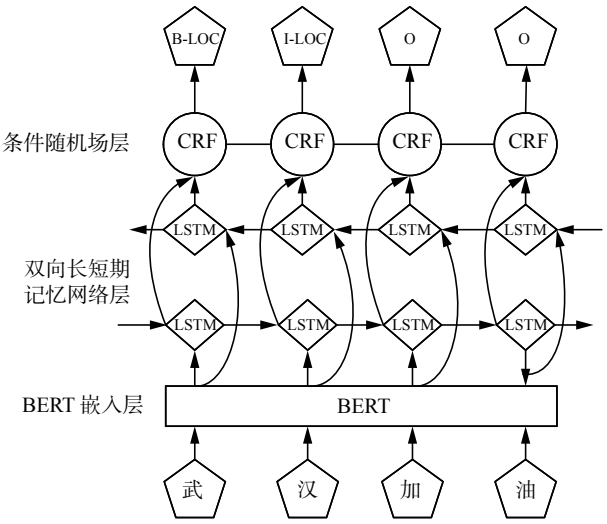

图 1　模型整体框架图

Fig. 1　Framework of model

BERT 的优点在其取得的成绩中显而易见，缺点是它在训练中消耗资源巨大，因此官方推荐使用微调的方式进行应用，即直接获取预训练好的模型，在实际应用模型训练中对 BERT 中的参数进行学习调整，此方法在一定程度上解决了计算资源消耗量大的问题。

在本文对 BERT 预训练模型的应用中，采用的是固定参数的方式。训练过程中不更新 BERT 内部参数，只对整体模型中除 BERT 以外的其他部分进行训练，虽然会损失一定的识别准确性，但是能够大幅减小模型训练过程中对计算机性能的压力，提高模型训练效率。

2.2　BiLSTM 层

长短期记忆网络在 1997 年被文献 [22] 提出，用以解决循环神经网络 (RNN) 容易梯度爆炸或梯度消失问题。此外它通过 3 个计算门的加入同时缓解了长序列遗忘问题，分别是遗忘门 f、输入门 i、输出门 o。

具体计算公式如下：

$$\begin{aligned} f_t &= \sigma(W_f h_{t-1} + U_f x_t + b_f) \\ i_t &= \sigma(W_i h_{t-1} + U_i x_t + b_i) \\ o_t &= \sigma(W_o h_{t-1} + U_o x_t + b_o) \\ \widetilde{c_t} &= \tanh(W_c h_{t-1} + U_c x_t + b_c) \\ c_t &= f_t \odot c_{t-1} + i_t \odot \widetilde{c_t} \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{1}$$

BiLSTM 是双向的长短期记忆网络，由一个前向 LSTM 及一个后向 LSTM 组成，分别计算，最后合并输出，使得模型不仅能学习到当前字的前向信息，同时能够学习到它的后向信息。BiLSTM 网络结构如图 2 所示。

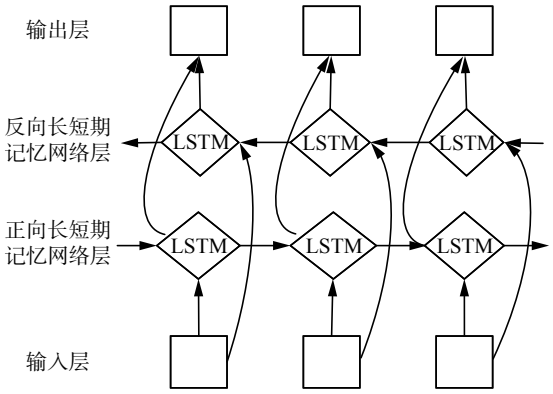

图 2　双向 LSTM 结构

Fig. 2　BiLSTM structure

2.3　CRF 层

条件随机场模型 (CRF) 是计算给定随机变量序列 $X=(X_1,X_2,\cdots,X_n)$ 的条件下，随机变量序列 $Y=(Y_1,Y_2,\cdots,Y_n)$ 的条件概率分布 $P(Y|X)$。模型假设随机变量序列满足马尔可夫性：

$$P(Y_i|X,Y_i,\cdots,Y_n)=P(Y_i|X,Y_{i-1},Y_{i+1}) \tag{2}$$

式中：X 表示输入观测序列；Y 表示对应的状态序列。条件随机场模型是一种无向概率图模型。2001 年 Lafferty 等[23] 提出了线性链条件随机场模型，线性链条件随机场是在序列标注任务中广泛应用的算法之一。

在序列标注任务中，一个字或词的标签与其相邻标签有一定的规则制约关系，例如 I 标签前面不会是 O 标签，I-LOC 不会跟在 B-PER 后面。CRF 能够通过学习得到各种标签序列的出现概率，通过概率选择减少不符合制约关系的标签序列出现。

设 $\boldsymbol{P}$ 为解码层输出的权重矩阵，进而可以得出评估分数 $S(x,y)$：

$$S(x,y)=\sum_{i=0}^{n} M_{y_i,y_{i+1}} + \sum_{i}^{n} P_{i,y_i} \tag{3}$$

式中：$\boldsymbol{M}$ 为转移矩阵；$M_{y_i,y_{i+1}}$ 表示从 y_i 标签转移到

y_{i+1} 标签的概率; P_{i,y_i} 表示第 i 个字被标记为 y_i 的概率; n 为序列长度。

最后采用极大似然法求解最大后验概率 $P(y|x)$, 获得模型的损失函数值。

$$\log P(y|x) = S(x,y) - \sum_{i=0}^{n} S(x,y_i) \tag{4}$$

2.4　加入自注意力层

为了弥补 BERT 不做微调训练所造成的模型命名实体识别准确性损失, 本文首先尝试在模型中增加自注意力 (self-attention) 层的方法。增加自注意力层模型如图 3 所示。

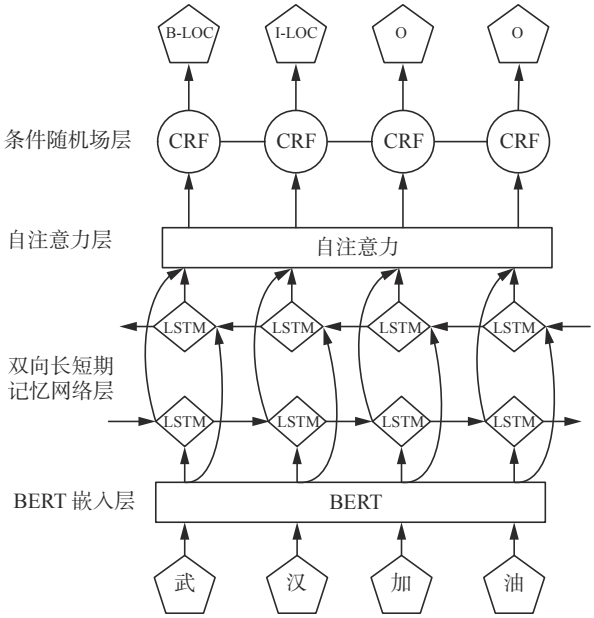

图 3　加入自注意力层的模型结构

Fig. 3　Model structure diagram with self-attention

文献 [24] 提出的自注意力机制相比于注意力 (attention) 机制, 自注意力只在序列内部完成注意力计算, 寻找序列内部联系, 常用放缩点积注意力进行计算, 计算公式为

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{SoftMax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

式中: $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 是来自同一输入与不同参数计算后得到的 3 个矩阵, 首先计算 $\boldsymbol{Q}$、$\boldsymbol{K}$ 矩阵乘法, 并除以 $\sqrt{d_k}$, 防止相乘结果过大, 最后利用 SoftMax 操作将结果归一化为概率分布, 并乘以矩阵 $\boldsymbol{V}$ 得到结果, 放缩点积自注意力结构如图 4 所示。此外, 为了能够多维度捕捉文本上下文特征, 本文使用了多头注意力机制。多头注意力机制是以不同参数多次重复计算 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 矩阵, 然后分别独立计算注意力, 将注意力计算结果拼接, 最后线性变换得到最终结果。计算方法为

$$\begin{aligned} &\text{head}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{Q}, \boldsymbol{K}\boldsymbol{W}_i^{k}, \boldsymbol{V}\boldsymbol{W}_i^{v}) \\ &\text{Multi}(\boldsymbol{Q},\boldsymbol{W},\boldsymbol{V}) = \text{Concat}(\text{head}_i \cdots \text{head}_h)\boldsymbol{W}^{o} \end{aligned} \tag{6}$$

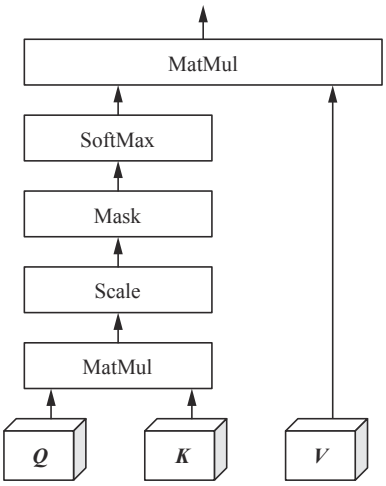

图 4　自注意力结构

Fig. 4　Self-attention structure

3　实验与分析

3.1　实验数据集与数据标注

本文使用两个数据集对提出模型的命名实体识别效果进行测试。一个是 SIGHAN 2006 竞赛 MSRA 数据集, 另一个是网络社交媒体 Weibo 数据集[25]。数据集详细信息如表 1 所示, 其中 MSRA 数据集不包含验证集。

表 1　数据集统计

Table 1　Statistics of datasets

数据集	类型	训练	验证	测试
MSRA	句子	46 364	—	4 365
	字	2 169 879	—	172 601
Weibo	句子	1 350	270	270
	字	73 778	14 509	14 842

命名实体主流标注模式分为 BIO 标注以及 BIOES 标注方式两种。本次实验所使用数据集均采用 BIO 标注, B 标签为命名实体首字, I 为命名实体非首字, O 为非命名实体。

MSRA 数据集包含地名 (LOC)、机构名 (ORG)、人名 (PER)3 类共 7 种标签。Weibo 数据集包含地名 (LOC)、机构名 (ORG)、人名 (PER) 和地理政治实体名 (GPE), 此外该数据集还以后缀的方式将命名实体细分为通用名 (NOM) 和专用名 (NAM) 两种, 其中地理政治类不包含通用名, 共 5 类实体 16 种标签。由于 Weibo 数据集可训练数据较少, 又因为取自网络社交媒体, 文字表达较为随意, 所以完成该数据集的命名实体识别任务更加困难。

3.2　实验环境及参数设置

本文实验模型选择采用 TensorFlow 进行搭建。实验硬件及软件环境配置如表 2 所示。

表 2 实验环境
Table 2 Experimental setting

项目	环境
内存	16 GB
GPU	NVIDIA GeForce RTX 2070
Python版本	Python 3.7.1
TensorFlow版本	TensorFlow 1.13.0

为了保证实验的公平性，各实验模型中除一些特别实验参数设置有所不同外，尽量保持参数的一致性。详细参数如下：单句最大长度限制 128 个字，LSTM 隐藏层维度为 200，dropout 为 0.5，使用 Adam 优化器，训练学习率 10^{-3}，MSRA 数据集 batch size 为 128，微博数据集 batch size 为 64。BERT 微调实验中训练学习率 5×10^{-5}，batch size 为 16。

3.3 实验结果与分析

实验 1 加入自注意力层对识别效果的影响。

本文首先提出在固定 BERT 参数的 BERT-BiLSTM-CRF 模型基础上，加入额外自注意力层的命名实体识别模型。并选取了 4 个与提出模型相关的深度神经网络模型与本文所提出的模型进行实验对比。4 种对照神经网络模型分别是：

1) BiLSTM-CRF，现阶段最常用的神经网络命名实体识别模型，由一个双向长短期记忆网络层和一个条件随机场层组成。

2) BiLSTM-self-attention-CRF 模型，在 BiLSTM-CRF 模型中加入了自注意力层的无预训练模型。

3) BERT 模型，直接利用数据集对 BERT 模型中的参数进行调整。

4) BERT-BiLSTM-CRF 模型，由 BERT 嵌入层，一个双向长短期记忆网络层以及一个条件随机场层所组成。

本文实验均采用精确率、召回率和 F1 值作为模型准确性的评价标准，计算公式如下：

$$\begin{cases} \mathrm{P}=\dfrac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\times100\% \\ \mathrm{R}=\dfrac{\mathrm{Tp}}{\mathrm{TP}+\mathrm{FN}}\times100\% \\ \mathrm{F1}=\dfrac{2\times\mathrm{P}\times\mathrm{R}}{\mathrm{P}+\mathrm{R}}\times100\% \end{cases} \tag{7}$$

式中：TP 表示正确识别出命名实体个数；FP 表示错误识别命名实体个数；FN 表示没有被识别命名实体个数；P 是精确率；R 是召回率。

表 3 是 5 种模型在 MSRA 数据集上的实验结果。

表 3 MSRA 数据集上的对比结果
Table 3 Experimental comparison results on MSRA

模型名称	P/%	R/%	F1/%	单轮时间/s
BiLSTM-CRF	89.88	87.93	88.89	42
BiLSTM-self-attention-CRF	90.52	87.44	88.95	46
BERT微调	94.91	94.72	94.82	1 320
BERT-BiLSTM-CRF	94.48	94.48	94.48	321
BERT-BiLSTM-self-attention-CRF	94.30	94.74	94.52	326

通过表 3 对比结果，可以看出，BERT 预训练模型的加入对模型命名实体识别准确性起到了明显提升作用，平均提升准确率 5.55%。但是 self-attention 层的加入对模型准确率提升效果有限。此外由于 BERT 微调实验对计算机性能要求较高，在本文的实验条件下，该模型的训练效率与其他模型相比有着明显的差距。

通过实验 1 说明 BERT 模型的加入对命名实体识别准确性有较大幅度的提升，但是训练所耗费的时间和对机器性能的要求大幅增加。固定 BERT 参数对模型准确性有一定影响，self-attention 层对模型识别准确性的提升效果不明显。

实验 2 减少 BERT 层数对识别效果带来的影响。

由于 BERT 模型每层在预训练中所学习到的信息不同，所以本文提出的第 2 个尝试性改进方案是减少 BERT 嵌入层数，从不同层获取模型输出结果，同样基于固定 BERT 参数的 BERT-BiLSTM-CRF 模型进行实验。首先在 MSRA 数据集上进行了 BERT 层数裁剪实验。分别取 BERT 模型嵌入层、第 3 层、第 6 层、第 9 层与 12 层进行对比，实验结果如表 4 所示。

表 4 在 MSRA 数据集上减少 BERT 模型层数的实验结果
Table 4 Experimental results of reducing the number of BERT model layers on the MSRA dataset

选取层	P/%	R/%	F1/%	单轮时间/s
嵌入层	90.51	90.17	90.34	43
3层	93.05	92.38	92.72	110
6层	94.07	94.49	94.28	188
9层	94.81	94.76	94.79	247
12层	94.48	94.48	94.48	321

实验结果显示，减小 BERT 嵌入层数，可以加快模型的训练速度。在嵌入层数小于 6 时，模型识别准确性随 BERT 层数的减少而下降，在嵌入

层数多于 6 时, 模型的识别准确性趋于平稳, 在嵌入 9 层时的实验结果优于 12 层 BERT 嵌入。表 5 给出了采用 9 层嵌入时, 模型对不同标签命名实体的识别结果。

表 5　本文模型对不同标签识别结果
Table 5　Different label recognition results

标签	P/%	R/%	F_1/%
地名LOC	96.33	94.91	95.62
机构名ORG	89.89	91.55	90.72
人名PER	96.40	97.43	96.91
全部ALL	94.81	94.76	94.79

在对 MSRA 数据集进行实验后, 为确保实验结果的普适性。本文继续在 Weibo 数据集上进行了实验。根据实验 1 的结果表明, 相对较深的层次嵌入效果较好, 所以本实验选取了第 8~12 层进行了多次实验, 实验结果如表 6 所示。

表 6　在 Weibo 数据集上减少 BERT 模型层数的实验结果
Table 6　Experimental results on reducing the number of BERT model layers on Weibo dataset

嵌入层数	P/%	R/%	F1/%	单轮时间/s
8	63.19	72.62	67.58	10
9	67.53	70.17	68.82	10
10	65.92	72.37	69.00	10
11	63.30	70.42	66.67	10
12	67.41	66.26	66.83	10

由于 Weibo 数据集数据量较少, 单轮时间对比相差不明显。在模型识别准确性上, 可以看出 9 层和 10 层的识别效果最好, 说明 BERT 模型在预训练中在第 9、10 层附近学习到的信息能够更好完成命名实体识别任务。

实验的最后, 本文对在 MSRA 数据集上模型识别错误的类型实例进行了总结, 常见错误如表 7 所示。

将所有包含实体识别错误的句子抽取并进行错误分析发现, 影响识别模型识别准确性的问题主要包括以下几类原因:

1) 实体标注模糊, 有些句子中“月 (指月亮)”是地点实体, 有些句子“月 (指月亮)”不是地点实体, 在同一个数据集中标注标准不统一, 如表 7 所列出的第 1 类。

2) 句子不完整或句子有错误导致句子无法被理解, 例如表 7 列出的第 2 类。

3) 实体嵌套, 在地名中包含人名, 机构名中包含地名, 机构名中包含机构名, 如表 7 所列出的第 3 类, 是模型在实体识别中出错概率最高的一类。

4) 特殊并列关系, 地名与人名并列, 地名与组织名并列, 如表 7 列出的第 4 类。

表 7　识别错误实例
Table 7　Identify error instances

分类	错误实例
第1类	原句1: 长安一片月, 万户捣衣声。 标注实体: 长安\月(LOC) 原句2: 明月几时有, 把酒问青天 标注实体: 无
第2类	原句: 某种香甜被太阳 标注实体: 太阳 识别实体: 无
第3类	原句: 东盟组织扩大 标注实体: 东盟 识别实体: 东盟组织
第4类	原句: 中国和东盟国家都发生了沧桑巨变 标注实体: 中国(LOC)\东盟(ORG) 识别实体: 中国(LOC)\东盟(LOC)

还有一种地名出现的地方如果填写人名句子通顺合理的复杂情况, 如果没有知识储备很难正确进行区分。

针对上述问题, 一方面需要明确统一语料标注标准才能通过高质量标注语料训练更好的命名实体自动识别模型, 还需要解决命名实体嵌套标注问题, 另一方面也侧面证明了类似 BERT 这一类海量预训练模型能够有效提高命名实体识别准确率的原因。

4　结束语

GPT、BERT 这类超大型预训练模型对自然语言处理研究领域的众多研究方向都带来了不小的提升, 但是它们巨大的资源消耗和时间消耗, 是不少研究者所承受不起的, 并且提高了领域的准入门槛和研究成本。ALBERT 的出现或许是一个可能, 但在本文之前的实验中其命名实体识别效果相比 BERT 有着较大差距。

本文首先针对 BERT 微调命名实体识别方法对计算机性能要求较高的问题, 采用了固定 BERT 参数的 BERT-BiLSTM-CRF 命名实体识别模型, 并尝试了两种方法对固定 BERT 参数的模型进行改进: 方法一向模型中继续添加自注意力层, 经过实验, 本方法并不能有效改善模型识别

效果；方法二通过缩减 BERT 模型嵌入层数对模型进行改进，经实验证明，该方法不仅能够大幅减小模型的训练时间，还能在一定程度上增强模型的实体识别效果。本文所改进的模型已经在电视台的创业投资栏目的智能机器人数据分析中得到初步应用。

本文方法在机构类实体的识别准确性上还有待提升，摆脱机构类实体嵌套这个实质性问题才能实现在准确性上进一步突破，未来的研究可以考虑采用阅读理解的方法，或通过改进标注形式来解决实体嵌套问题。另外通过实验最后的错误实例分析可以看出，制定完善的实体标注标准也是提高实体识别效果的重要保障。本文仅在命名实体识别任务上对减少 BERT 嵌入层数对模型识别效果改善作用进行了验证，是否在其他序列标注任务中有同样的结论是下一步研究的目标。

参考文献：

[1] 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.
LIU Liu, WANG Dongbo. A review on named entity recognition[J]. Journal of the China society for scientific and technical information, 2018, 37(3): 329–340.

[2] BIKEL D M. An algorithm that learns what's in a name[J]. Machine learning, 1999, 34(1/2/3): 211–231.

[3] MAYFIELD J, MCNAMEE P, PIATKO C D, et al. Named entity recognition using hundreds of thousands of features[C]//North American Chapter of the Association for Computational Linguistics. Edmonton, Canada, 2003: 184−187.

[4] MCCALLUM A, LI W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]//North American Chapter of the Association for Computational Linguistics. Edmonton, Canada, 2003: 188−191.

[5] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011, 12(1): 2493–2537.

[6] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL].[2015-08-09]. https://arxiv.org/abs/1508.01991.

[7] 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, Germany, 2016: 1064−1074.

[8] 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.

[9] YANG Y, CHEN W, LI Z, et al. Distantly supervised NER with partial annotation learning and reinforcement learning[C]//International Conference on Computational Linguistics. Santa Fe, USA, 2018: 2159−2169.

[10] 彭嘉毅, 方勇, 黄诚, 等. 基于深度主动学习的信息安全领域命名实体识别研究 [J]. 四川大学学报(自然科学版), 2019, 56(3): 457–462.
PENG Jiayi, FANG Yong, HUANG Cheng, et al. Cyber security named entity recognition based on deep active learning[J]. Journal of sichuan university (natural science edition), 2019, 56(3): 457–462.

[11] 朱艳辉, 李飞, 冀相冰, 等. 反馈式 K 近邻语义迁移学习的领域命名实体识别 [J]. 智能系统学报, 2019, 14(4): 820–830.
ZHU Yanhui, LI Fei, JI Xiangbing, et al. Domain-named entity recognition based on feedback k-nearest semantic transfer learning[J]. CAAI transactions on intelligent systems, 2019, 14(4): 820–830.

[12] 王红斌, 沈强, 线岩团. 融合迁移学习的中文命名实体识别 [J]. 小型微型计算机系统, 2017, 38(2): 346–351.
WANG Hongbin, SHEN Qiang, XIAN Yantuan. Research on Chinese named entity recognition fusing transfer learning[J]. Journal of Chinese computer systems, 2017, 38(2): 346–351.

[13] 冯鸾鸾, 李军辉, 李培峰, 等. 面向国防科技领域的技术和术语识别方法研究 [J]. 计算机科学, 2019, 46(12): 231–236.
FENG Luanluan, LI Junhui, LI Peifeng, et al. Technology and terminology detection oriented national defense science[J]. Computer science, 2019, 46(12): 231–236.

[14] 杨维, 孙德艳, 张晓慧, 等. 面向电力智能问答系统的命名实体识别算法 [J]. 计算机工程与设计, 2019, 40(12): 3625–3630.
YANG Wei, SUN Deyan, ZHANG Xiaohui, et al. Named entity recognition for intelligent answer system in power service[J]. Computer engineering and design, 2019, 40(12): 3625–3630.

[15] 李冬梅, 檀稳. 植物属性文本的命名实体识别方法研究 [J]. 计算机科学与探索, 2019, 13(12): 2085–2093.
LI Dongmei, TAN Wen. Research on named entity recognition method in plant attribute text[J]. Journal of frontiers of computer science and technology, 2019, 13(12): 2085–2093.

[16] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]//Annual meeting of the association for computational linguistics. Melbourne, Australia, 2018: 1554−1564.

[17] 李明扬, 孔芳. 融入自注意力机制的社交媒体命名实体

识别 [J]. 清华大学学报(自然科学版), 2019, 59(6): 461–467.
LI Mingyang, KONG Fang. Combined self-attention mechanism for named entity recognition in social media[J]. Journal of Tsinghua university (science and technology edition), 2019, 59(6): 461–467.

[18] CAO P, CHEN Y, LIU K, et al. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism[C]//Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 182−192.

[19] PETERS M E, RUDER S, SMITH N A, et al. To tune or not to tune? Adapting pretrained representations to diverse tasks.[C]//Proceedings of the 4th Workshop on Representation Llearning for NLP. Florence, Italy, 2019: 7−14.

[20] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. Computation and language, 2018(10): 1810–4805.

[21] YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding[C]// Neural Information Processing Systems. Vancouver, Canada, 2019: 5753−5763.

[22] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735–1780.

[23] LAFFERTY J, MCCALLUM A, PEREIRA F, et al. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//International Conference on Machine Learning. San Francisco, USA, 2001: 282−289.

[24] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 5998−6008.

[25] HE H, SUN X. F-Score driven max margin neural network for named entity recognition in Chinese social media[C]// Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain, 2017: 713−718.

作者简介:

毛明毅, 副教授, 博士, 中国人工智能学会高级会员, 主要研究方向为人工智能基础理论、泛逻辑学, 主持和参与国家自然基金项目和北京市自然科学基金项目及其他纵向课题 8 项, 主持横向课题 10 余项, 获专利授权和软件著作权 10 余项, 获得全国竞赛“优秀指导教师”等多种荣誉。发表学术论文 50 余篇, 出版专著 2 部。

吴晨, 硕士研究生, 主要研究方向为人工智能基础、智能机器人、自然语言理解。

钟义信, 教授, 博士生导师, 发展中世界工程技术科学院院士, 中国人工智能学会原理事长, 现任国际信息研究学会中国分会主席, 北京邮电大学−格分维人工智能联合实验室学术委员会主任, 主要研究方向为通信理论、信息科学、人工智能。主持国家级和省部级项目数十项。先后提出和建立“全信息理论”“全信息自然语言理解理论”“机制主义人工智能统一理论”以及“机器知行学”理论, 发现和总结了“信息转换与智能创生定律”, 先后获得“有突出贡献的归国留学人员”、“全国优秀教师”等称号; 获得首届吴文俊科学技术成就奖和首届中国电子学会信息理论杰出贡献奖。发表学术论文 500 余篇, 出版学术专著 18 部。