Computer Science and Application 计算机科学与应用, 2020, 10(5), 883-892
Published Online May 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.105091

Hans汉斯

Chinese Named Entity Recognition Method Based on ALBERT

Boyan Deng, Lianglun Cheng

School of Computers, GDUT, Guangzhou Guangdong
Email: 864811548@qq.com, llcheng@gdut.edu.cn

Received: Apr. 20th, 2020; accepted: May 5th, 2020; published: May 12th, 2020

Abstract

The BERT pre-trained language model has been widely used in Chinese named entity recognition due to its good performance, but the large number of parameters and long training time has limited its practical application scenarios. In order to solve these problems, we propose ALBERT-BiLSTM-CRF, a model for Chinese named entity recognition task based on ALBERT. Structurally, the model firstly trains character-level word embeddings on large-scale text through the ALBERT pre-training language model, and then inputs the word embeddings into the BiLSTM model to obtain more inter-character dependencies, and finally decodes through CRF and extracts the corresponding entities. This model combines the advantages of ALBERT and BiLSTM-CRF models to identify Chinese entities, and achieves an F1 value of 95.22% on the MSRA dataset. Experiments show that while greatly reducing the pre-training parameters, the model retains relatively good performance and has good scalability.

Keywords

Chinese Named Entity Recognition, ALBERT, Pre-Trained Language Model, BiLSTM, CRF

基于ALBERT的中文命名实体识别方法

邓博研，程良伦

广东工业大学，广东 广州
Email: 864811548@qq.com, llcheng@gdut.edu.cn

收稿日期：2020年4月20日；录用日期：2020年5月5日；发布日期：2020年5月12日

摘　要

在中文命名实体识别任务中，BERT预训练语言模型因其良好的性能得到了广泛的应用，但由于参数量

文章引用: 邓博研, 程良伦. 基于 ALBERT 的中文命名实体识别方法[J]. 计算机科学与应用, 2020, 10(5): 883-892.
DOI: 10.12677/csa.2020.105091

过大、训练时间长，其实际应用场景受限。针对这个问题，提出了一种基于ALBERT的中文命名实体识别模型ALBERT-BiLSTM-CRF。在结构上，先通过ALBERT预训练语言模型在大规模文本上训练字符级别的词嵌入，然后将其输入BiLSTM模型以获取更多的字符间依赖，最后通过CRF进行解码并提取出相应实体。该模型结合ALBERT与BiLSTM-CRF模型的优势对中文实体进行识别，在MSRA数据集上达到了95.22%的F1值。实验表明，在大幅削减预训练参数的同时，该模型保留了相对良好的性能，并具有很好的可扩展性。

关键词

中文命名实体识别，ALBERT，预训练语言模型，BiLSTM，条件随机场

1. 引言

命名实体识别(Named Entity Recognition, NER)任务一直是自然语言处理(Natural Language Processing, NLP)领域中的研究热点，主要采用序列标注的方式，获取文本中具有特定意义的实体。从早期基于规则和字典的方法开始，历经传统机器学习方法和深度学习方法，到近两年来的预训练语言模型，研究者们进行了不断的尝试与探索。

在 NER 任务中，传统机器学习方法，如隐马尔可夫模型(Hidden Markov Model, HMM)、最大熵马尔可夫模型(Maximum Entropy Markov Models, MEMM)和条件随机场(Conditional Random Field, CRF)，利用概率图模型的思想，计算当前序列中的状态转移概率，进而获取最优标注序列[1]。另一方面，深度学习方法，如循环神经网络(Recurrent Neural Network, RNN)，逐渐演化出长短期记忆网络(Long Short-Term Memory, LSTM)与门控循环单元(Gated Recurrent Unit, GRU)等具有记忆功能的网络结构，通过融合更多信息获取更好的效果[2]。2013 年，Mikolov 等人提出词向量化工具 Word2Vec [3]，如何训练词嵌入(Word Embedding)成为新一轮的研究热点。2018 年，Devlin 等人结合前人的经验，提出了 BERT (Bidirectional Encoder Representations from Transformers)预训练语言模型[4]，在 NER 等 11 个 NLP 任务上取得了最好效果。为解决 BERT 参数量过大的问题，Lan 等人于 2019 年提出了 ALBERT (A Lite BERT)预训练语言模型[5]，在基本维持性能的前提下，获得了更好的模型扩展性。

本文将 ALBERT 预训练语言模型与 BERT 预训练语言模型在 MSRA 公开的中文命名实体识别数据集上进行对比，并结合 BiLSTM-CRF 模型，构建 ALBERT-BiLSTM-CRF 模型，在数据集上进行了进一步实验验证，结果表明 ALBERT 预训练语言模型保留了相对良好的性能，并具有很好的可扩展性。

2. 相关工作

与英文不同，中文构成较为复杂，不仅在语义上存在字符与词语的区分，还有笔划与部首等额外信息。Dong 等人应用融合了字符嵌入和部首级表示的 BiLSTM-CRF 模型[6]，在没有精心调整特征的情况下取得了更好的效果。Xiang 等人提出了一种字词混合嵌入(CWME)方法[7]，可以结合下游的神经网络模型有效地提高性能。Zhang 等人提出了 Lattice LSTM [8]，对输入字符序列以及与词典匹配的所有潜在词进行编码，显式地利用了字符信息和词序信息，避免了分词错误，并在 MSRA 公开的中文命名实体识别数据集上达到了 93.18%的 F1 值。针对微博数据集等口语化较多的语料，Xu 等人提出了 ME-CNER [9]，

以推导中文文本中命名实体的字符嵌入，在微博数据集上实现了较大的性能改进。Johnson 等人提出了一种综合嵌入方法，CWPC_BiAtt [10]，将字符嵌入、词语嵌入和词性嵌入依照顺序进行拼接以获得它们之间的依存关系，并采用注意力机制捕获当前位置内容与任何位置之间的联系，在 MSRA 数据集和微博 NER 语料库上获得了较高的精度和召回率。

以上方法的探索证明了字符嵌入与词语嵌入在中文命名实体识别任务中的有效性，而在 Devlin 等人提出 BERT 后，其输出的词嵌入所具有的良好的语义表达效果，使针对预训练语言模型的研究成为热点。在中文电子病历的命名实体识别任务中，Dai 等人[11]与 Jiang 等人[12]使用 BERT-BiLSTM-CRF 模型获得了优于其他模型的效果。Cai 等人[13]先通过 BERT 预训练语言模型增强字符的语义表示，然后将字符嵌入输入 BiGRU-CRF 进行训练，最终得到了优于最佳模型的效果。Gong 等人[14]使用 BERT 训练汉字嵌入，将其与汉字根基表示法联系起来，并将结果放入 BiGRU-CRF 模型中，在中文数据集上取得了良好的效果。为进一步提升 BERT 在中文 NER 任务上的表现，Cui 等人提出了全字掩码(WWM) [15]，实验表明该方法可以带来显著收益。

尽管 BERT 具有非常优异的性能表现，但由于其庞大的参数量，很多研究者开始研究如何在保持一定性能的情况下压缩 BERT 的规模。Michel 等人发现，在测试时移除大量注意力表头也不会对性能产生显著影响[16]。Wang 等人提出了一种基于低秩矩阵分解与强化的拉格朗日 L0 范数正则化的新型结构化修剪方法[17]，在任何稀疏度级别上都可以显著实现推理加速。Shen 等人提出了一种新的逐组量化方案，并使用基于 Hessian 的混合精度方法进一步压缩模型[18]。Lan 等人[5]提出的 ALBERT 采用两种参数减少技术以降低内存消耗并提高 BERT 的训练速度，同时使用了一种自我监督的损失以提高训练效果。

Lan 等人提出的 ALBERT 具有很好的减少参数的效果，但参数的减少必然带来性能的损失。为提升下游任务的性能，本文提出结合 BiLSTM-CRF 模型的 ALBERT-BiLSTM-CRF 模型，并在 MSRA 公开的中文命名实体识别数据集上达到了 95.22%的 F1 值。

3. ALBERT-BiLSTM-CRF 模型

模型主要由三部分构成，分别为 ALBERT 预训练语言模型、BiLSTM 层和 CRF 层。以 ALBERT 的编码输出作为 BiLSTM 层的输入，再在 BiLSTM 的隐藏层后加一层 CRF 层用以解码，最终得到每个字符的标注类型。具体结构如图 1。

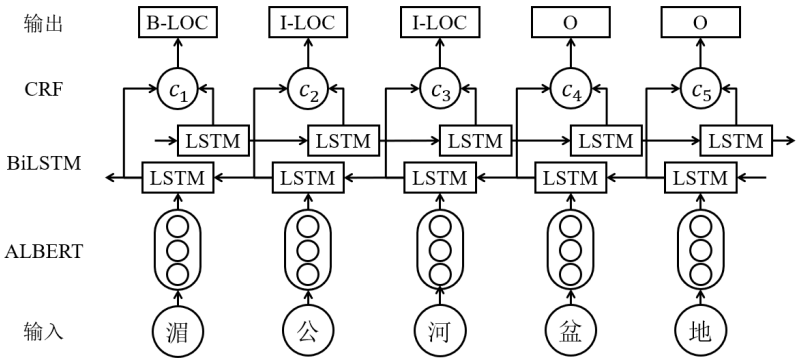

Figure 1. ALBERT-BiLSTM-CRF Model
图 1. ALBERT-BiLSTM-CRF 模型

3.1. BERT 预训练语言模型

在自然语言处理任务中，语言模型是一个重要概念，其任务是计算语言序列 $w_1, w_2, \cdots, w_n$ 的出现概率

$p(w_1, w_2, \cdots, w_n)$：

$$p(S) = p(w_1, w_2, \cdots, w_n) = \prod_{i=1}^{n} p(w \mid w_1, w_2, \cdots, w_{i-1}) \tag{1}$$

传统语言模型，如神经网络语言模型，一方面由于其是单向的，无法融入上下文信息，另一方面由于训练得到的词嵌入为固定的，无法表示词的多义性。BERT 的模型结构很好地解决了这两个问题。

BERT 模型结构如图 2 所示。相比于 GPT [19]和 ELMO [20]，BERT 结合了二者的长处，采用双向 Transformer [21]作为编码器，一方面使用效果更好的 Transformer 替代了 LSTM，另一方面，双向的语言模型使得 BERT 可以获取上下文信息，进而使词嵌入具有更丰富的语义信息。

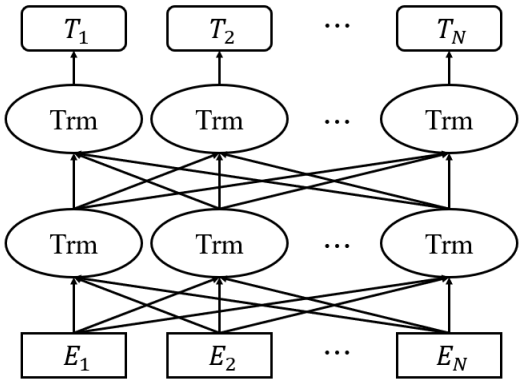

Figure 2. BERT Pre-trained Language Model
图 2. BERT 预训练语言模型

Transformer 编码单元结构如图 3 所示。Transformer 主要利用的思想是注意力机制，通过自注意力获取序列内部联系：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{2}$$

然后通过多头结构拼接结果：

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \cdots, \text{head}_h)W^O \tag{3}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{4}$$

而在 Transformer 编码单元中，为使网络更容易训练，引入了残差连接和层归一化：

$$LN(x_i) = \alpha \times \frac{x_i - \mu_L}{\sqrt{\sigma_L^2 + \varepsilon}} + \beta \tag{5}$$

$$FFN = \max(0, xW_1 + b_1)W_2 + b_2 \tag{6}$$

由于注意力机制本身不提取时序特征，在输入端通过添加位置嵌入引入位置信息：

$$PE_{(pos,2i)} = \sin\left(pos / 10000^{2i/d_{\text{model}}}\right) \tag{7}$$

$$PE_{(pos,2i+1)} = \cos\left(pos / 10000^{2i/d_{\text{model}}}\right) \tag{8}$$

最后，BERT 将位置嵌入和词嵌入拼接起来作为模型输入。

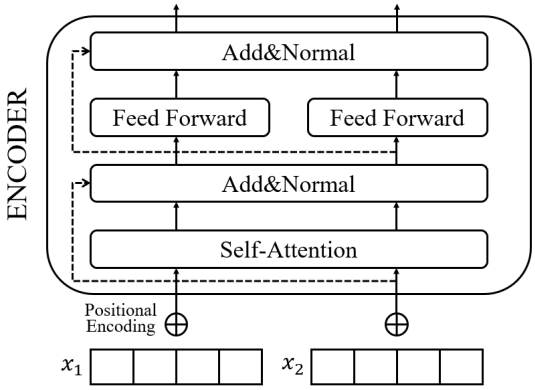

Figure 3. Transformer Encoder
图 3. Transformer 编码单元

虽然在模型结构上只是对 GPT 和 ELMO 的一种改进，但 BERT 创新性地提出了“Masked 语言模型”和“下一句预测”两种任务。Masked 语言模型随机选择语料中 15%的词用[Mask]掩码代替，而这些被选择的词中有 80%被正常替换，10%被替换成另外一个词，10%保持不变；下一句预测是在做语言模型预训练的时候，正样本选择同文档中顺序相连的两个句子，负样本从不同文档中随机选择句子拼到第一个句子后面。这两种任务，前者获取词间信息，后者获取句子间信息，在训练的时候融入这两种信息可以使词嵌入具有更好的全局表达效果。

3.2. ALBERT 模型的改进

相比于 BERT，ALBERT 主要采用了三种改进方法。

3.2.1. 对嵌入的因式分解

在 BERT 中，词嵌入大小 E 与隐藏层大小 H 相同，总参数量为词汇表长度 V 乘以每个词嵌入隐藏层大小 H，其复杂度为 $\mathrm{O}(V\times H)$。而 ALBERT 认为，词级别的嵌入没有上下文依赖的表述，而隐藏层的输出包括了一些上下文信息，理论上来说隐藏层的表述包含的信息应该更多，应该让 $H \gg E$。因此 ALBERT 提出利用因式分解降低参数的方法，先把 one-hot 向量映射到一个大小为 E 的低维度空间，然后通过高维映射变换到隐藏层空间内，进而使参数量的复杂度变换如下：

$$\mathrm{O}(V\times H)\to \mathrm{O}(V\times E+E\times H) \tag{9}$$

3.2.2. 跨层参数共享

Transformer 共享参数有多种方案，ALBERT 采用了全连接层与注意力层都进行参数共享的策略，共享了编码器内的所有参数，即用一个自注意力层的参数共享给每个注意力头。由于全连接层与注意力层占据了总参数量中的大部分，而跨层参数共享策略直接把二者的参数减少为单头时的情况，极大地压缩了参数总量。另一方面，减少了模型复杂度的同时，也使性能有比较明显的降低。

3.2.3. 句间连贯性损失

为了弥补参数减少带来的性能损失，ALBERT 采用了一系列的方法提升模型的性能，例如放宽输入序列限制，采用 n-gram 掩码代替单字掩码，移除 dropout，而性能提升最明显的是一种新的训练任务，句间连贯性损失。ALBERT 认为，相比于 Masked 语言模型的提升，BERT 采用的下一句预测任务降低了下游任务的性能，因为该任务包含了两个子任务，主题预测与关系一致性预测，而前者比后者简单很多。在 ALBERT 中，采用了一种新的策略对下一句预测任务进行改进，仅保留关系一致性预测。该任务中，

正样本与下一句预测任务相同，选择同文档中两个顺序相连的句子，负样本则由正样本的两个句子调换顺序获得。

3.3. BiLSTM 层

对于文本序列的时序特征，神经网络模型中的 RNN 具有很好的效果。RNN 由输入层 x、隐藏层 h 和输出层 y 组成，在命名实体标注任务的上下文中，x 表示输入的特征，y 表示输出的标签。与前馈神经网络相比，RNN 在隐藏层引入了上个隐藏状态与当前隐藏状态的连接，计算如下：

$$h(t)=f\left(\mathrm{U}x(t)+\mathrm{W}h(t-1)\right) \tag{10}$$

其中，U 与 W 是在训练时计算的连接权重，$f(z)$ 是 Sigmoid 激活函数。

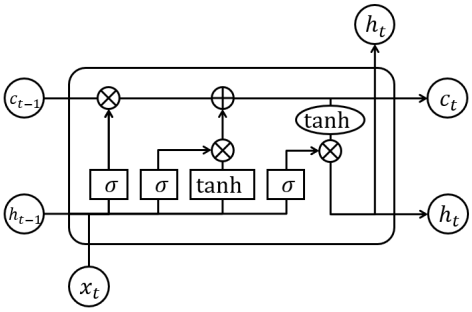

Figure 4. Cell State in LSTM
图 4. LSTM 中的细胞状态结构

为了更好地获取文本中的长范围依赖信息，LSTM 在隐藏层用专门构建的记忆细胞结构代替了 RNN 中的更新模块，如图 4。其主要组成部分为输入门 i、遗忘门 f、输出门 o 和记忆细胞 c，构造如下：

$$i_t=\sigma\left(W_{xi}x_t+W_{hi}h_{t-1}+W_{ci}c_{t-1}+b_i\right) \tag{11}$$

$$f_t=\sigma\left(W_{xf}x_t+W_{hf}h_{t-1}+W_{cf}c_{t-1}+b_f\right) \tag{12}$$

$$c_t=f_tc_{t-1}+i_t\tanh\left(W_{xc}x_t+W_{hc}h_{t-1}+b_c\right) \tag{13}$$

$$o_t=\sigma\left(W_{xo}x_t+W_{ho}h_{t-1}+W_{co}c_t+b_o\right) \tag{14}$$

隐藏层 h 的更新由 o 与 c 共同决定：

$$h_t=o_t\tanh\left(c_t\right) \tag{15}$$

BiLSTM 在 LSTM 的基础上引入反向传递的思想，构建了两个方向相反的隐藏层。通过获取上下文信息，BiLSTM 可以取得更好的语义表达效果。

3.4. CRF 层

当前命名实体识别任务的解决方案实质上是一种序列标注任务，而 CRF 是神经网络模型之外的一种主流模型。与神经网络模型相比，CRF 可解释性强，每一个特征都有可以具体解释的意义，而且 CRF 可以捕捉标签之间的依赖关系，通过概率转移矩阵排除非法用语的情况。

对于给定输入序列 $x=\left(x_1,x_2,\cdots,x_n\right)$ 和对应的标签序列 $y=\left(y_1,y_2,\cdots,y_n\right)$，CRF 定义评估分数为：

$$s(x,y)=\sum_{i=1}^{n}W_{y_{i-1},y_i}+\sum_{i=1}^{n}P_{i,y_i} \tag{16}$$

其中，W 是转移矩阵，W_{y_{i-1},y_i} 表示从标签 y_{i-1} 到标签 y_i 的转移概率，P_{i,y_i} 表示输入 x_i 映射到标签 y_i 的非归一化概率。

输入序列到标签序列的对应概率 $p(y|x)$ 可用 Softmax 函数进行计算：

$$p(y|x)=\frac{e^{s(x,y)}}{\sum_{\tilde{y}\in Y_x} e^{s(x,\tilde{y})}} \tag{17}$$

其中，Y_x 为所有可能预测的标签序列。

在训练中，最大化 $p(y|x)$ 的对数似然，可将损失函数定义为：

$$-\log\left(p(y|x)\right)=\log\left(\sum_{\tilde{y}\in Y_x} e^{s(x,\tilde{y})}\right)-s(x,y) \tag{18}$$

解码时，选择 y^* 作为输出预测标签序列，通过动态规划算法求得最优解：

$$y^*=\arg\max_{\tilde{y}\in Y_x} s(x,\tilde{y}) \tag{19}$$

4. 实验及结果分析

4.1. 实验数据

本文使用微软亚洲研究院(MSRA)公开的中文命名实体识别数据集。该数据集包含人名(PER)、地名(LOC)与机构名(ORG)三类实体，以约 12:1 的比例划分训练集和测试集，具体实体个数统计如表 1。

Table 1. Statistics of the number of entities
表 1. 实体个数统计

数据集	人名	地名	机构名	共计
训练集	17615	36517	20571	74703
测试集	1973	2877	1331	6181

4.2. 标注策略与评价指标

本文使用的标注策略为 BIO 标注模式，即以 B-X 标记实体起始，以 I-X 标记实体中间，以 O 标记无关字符。MSRA 数据集包含人名、地名与机构名三类实体，X 可分别取“PER”、“LOC”与“ORG”，因此待预测标签共有 7 种，分别为“B-PER”，“I-PER”，“B-LOC”，“I-LOC”，“B-ORG”，“I-ORG”和“O”。

本文使用的评价指标为精确率(P)，召回率(R)和 $F1$ 值，具体定义如下：

$$P=\frac{T_P}{T_P+F_P}\times 100\% \tag{20}$$

$$R=\frac{T_P}{T_P+F_N}\times 100\% \tag{21}$$

$$F1=\frac{2PR}{P+R}\times 100\% \tag{22}$$

其中，T_P 为模型正确识别到相关实体的样本数，F_P 为模型将非相关实体误识别为相关实体的样本数，F_N 为模型未识别到其中所包含的相关实体的样本数。

4.3. 实验过程

为验证模型效果，本文使用以下模型进行对比：

Lattice-LSTM-CRF 模型[8]，由 Zhang 等人于 2018 年提出，融合了字符信息和词序信息，在 MSRA 上达到了当时最好的命名实体识别效果。

Bert-base 模型[4]，由 Devlin 等人于 2018 年提出的预训练语言模型。其预训练的结果进行 fine-tuning 之后，在当时的各项自然语言处理任务中均获得了最好的效果。BERT-base 的 fine-tuning 效果在本文实验中作为命名实体识别任务的基准。

ALBERT-base 模型[5]，由 Lan 等人于 2019 年提出的预训练语言模型。ALBERT 的实质是进行了参数削减的轻量版 BERT，相比 BERT 的优势在于更少的参数占用更少的内存，并可进行更快的训练。

BERT-BiLSTM-CRF 模型，Dai 等人[11]与 Jiang 等人[12]所使用的当前主流 NER 模型，很多研究者将此模型作为参照，与自己所提出的模型进行对比，以寻求进一步提升 NER 效果的方法。

4.4. 实验环境与参数设置

本文进行实验时所采用的环境如表 2。

Table 2. Statistics of the number of entities
表 2. 实验环境

配置	规格
操作系统	Ubuntu
CPU	i7-7700@3.6GHz
GPU	GTX 2080(8GB)
Python	3.7.3
Tensorflow	1.13.1
内存	16GB DDR4

实验参数方面，本文主要对模型规格选择与训练参数做出设置。预训练语言模型选择中，BERT 采用 BERT-base，使用 BERT 中文预训练模型文件 chinese_L-12_H-768_A-12 作为预训练模型；ALBERT 采用 ALBERT-base，使用参数量 12M，层数 12，大小为 40M 的中文预训练模型。训练参数选择上，最大序列长度为 128，训练批次大小为 64，学习率为 2e-5，丢弃率为 0.1。

4.5. 实验结果及分析

ALBERT-BiLSTM-CRF 模型在人名、地名和机构名三类实体上的精确率、召回率与 F1 值如表 3。

Table 3. Statistics of the number of entities
表 3. 不同类型命名实体的识别效果

类型	*P*	*R*	*F*1
PER	97.87	96.40	97.13
LOC	97.36	90.72	93.92
ORG	96.69	93.61	95.13

从表中可以看出，ALBERT-BiLSTM-CRF 模型在三类实体上的精确率都比较高，但在地名和机构名上都有明显偏低的召回率。经过对数据集的分析发现，地名与机构名这两类实体存在部分漏标、误标的

情况。如在测试集中将“江苏省公安部[ORG]”标记为“江苏省[LOC]公安部”，这将导致其中的“江苏省[LOC]”无法识别，进而降低地名类实体的召回率。而“工商联”、“理事会”等机构名类实体在训练集中存在漏标的情况，这将导致模型不能很好地识别这些实体，进而降低其精确率与召回率。

为进一步验证模型性能，将本文提出的模型与其他模型在数据集上进行对比，效果如表 4。

Table 4. Statistics of the number of entities
表 4. 不同模型的命名实体识别效果

模型	*P*	*R*	*F*1
Lattice-LSTM-CRF	93.57	92.79	93.18
BERT-fine-tuning	96.68	96.01	96.33
ALBERT-fine-tuning	95.15	94.11	94.62
BERT-BiLSTM-CRF	97.92	96.88	97.40
ALBERT-BiLSTM-CRF	97.38	93.16	95.22

从表 4 中可以看出，Lattice-LSTM-CRF 模型性能低于作为基准的 BERT-base 的 fine-tuning 效果，这说明单纯的神经网络模型即使加入字词融合特征等特别设计的语义结构信息，在效果上仍旧不如预训练语言模型。另一方面，该模型在性能上全面低于基于预训练语言模型的方法，说明预训练语言模型抽取语义信息的能力比较强，其在海量文本中抽取到的一般特征比手工设计的结构特征要好。

在预训练语言模型之间进行比较可以发现，基于 ALBERT 的模型相比于基于 BERT 的模型整体表现低 1%~2%，而 BiLSTM-CRF 对于二者的整体性能提升基本相似。值得注意的是，ALBERT-BiLSTM-CRF 在召回率上相比于仅对 ALBERT 进行微调的效果反而低了 0.9%，这可能是因为 ALBERT 放弃了 dropout 策略，而 BiLSTM 在模型中保留了 dropout 的操作，因此在语义表达上产生了些微冲突。

本文在实验中使用的两种预训练语言模型都是 base 规模，而在模型结构相同时，参数量少很多的 ALBERT 自然会丢失一部分性能。另外，参数的减少使训练效率得到了提高，实验过程中基于 ALBERT 的模型的训练速度比基于 BERT 的略快，且占用内存较少。

5. 结束语

为解决 BERT 预训练语言模型参数量过大且训练时间长的问题，本文提出了 ALBERT-BiLSTM-CRF 模型，使用 ALBERT 预训练语言模型进行词嵌入的训练，并在下游任务采用 BiLSTM-CRF 模型以进一步提升中文命名实体识别任务的效果。该模型能够利用 BiLSTM-CRF 的模型特性充分学习字符间的依赖信息，而用 ALBERT 进行预训练既保留了 BERT 在海量文本上充分学习文本特征信息的优点，又通过参数削减提高了预训练语言模型的扩展能力，使 large 等更大规模的预训练语言模型在性能受限的机器上的应用成为了可能。

通过在 MSRA 公开数据集上进行验证，本文所提出的 ALBERT-BiLSTM-CRF 模型在大幅降低预训练语言模型参数的同时保留了相对良好的性能，这也同时验证了 ALBERT 的扩展潜力，可为后续研究带来一定的参考价值。考虑到 BiLSTM 模型本身参数量不小，下一步研究主要针对 ALBERT 本身的优化方向，可尝试一些新的训练任务或训练策略以进一步提高任务表现。

致　谢

感谢以下基金项目的支持与帮助：国家自然科学基金广东联合基金，离散制造过程人工智能驱动的优化与控制，项目编号 U1801263；工业过程数据实时获取与知识自动化，NSFC-广东联合基金，项目编

号 U1701262；广东省信息物理融合系统重点实验室，项目编号：2016B030301008。

参考文献

[1] Lafferty, J., McCallum, A. and Pereira, F.C.N. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data.

[2] Jozefowicz, R., Zaremba, W. and Sutskever, I. (2015) An Empirical Exploration of Recurrent Network Architectures. *Proceedings of the* 32*nd International Conference on Machine Learning*, **37**, 2342-2350.

[3] Mikolov, T., Chen, K., Corrado, G., *et al*. (2013) Efficient Estimation of Word Representations in Vector Space. arXiv Preprint arXiv:1301.3781.

[4] Devlin, J., Chang, M.W., Lee, K., *et al*. (2018) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv Preprint arXiv:1810.04805.

[5] Lan, Z., Chen, M., Goodman, S., *et al*. (2019) Albert: A Lite Bert for Self-Supervised Learning of Language Representations. arXiv Preprint arXiv:1909.11942.

[6] Dong, C., Zhang, J., Zong, C., *et al*. (2016) Character-Based LSTM-CRF with Radical-Level Features for Chinese Named Entity Recognition. In: Lin, C.-Y., Xue, N.W., Zhao, D.Y., Huang, X.J. and Feng, Y.S., Eds., *Natural Language Understanding and Intelligent Applications*, Springer, Cham, 239-250. https://doi.org/10.1007/978-3-319-50496-4_20

[7] Xiang, Y. (2017) Chinese Named Entity Recognition with Character-Word Mixed Embedding. *Proceedings of the* 2017 *ACM on Conference on Information and Knowledge Management*, Singapore, November 2017: 2055-2058.

[8] Zhang, Y. and Yang, J. (2018) Chinese NER Using Lattice LSTM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne. https://doi.org/10.18653/v1/P18-1144

[9] Xu, C., Wang, F., Han, J., *et al*. (2019) Exploiting Multiple Embeddings for Chinese Named Entity Recognition. *The* 28*th ACM International Conference on Information and Knowledge Management*, Beijing, November 2019, 2269-2272. https://doi.org/10.1145/3357384.3358117

[10] Johnson, S., Shen, S. and Liu, Y. (2020) CWPC_BiAtt: Character-Word-Position Combined BiLSTM-Attention for Chinese Named Entity Recognition. *Information*, **11**, 45. https://doi.org/10.3390/info11010045

[11] Dai, Z., Wang, X., Ni, P., *et al*. (2019) Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records. 2019 12*th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics* (*CISP-BMEI*), Suzhou, 19-21 October 2019, 1-5. https://doi.org/10.1109/CISP-BMEI48845.2019.8965823

[12] Jiang, S., Zhao, S., Hou, K., *et al*. (2019) A BERT-BiLSTM-CRF Model for Chinese Electronic Medical Records Named Entity Recognition. 2019 12*th International Conference on Intelligent Computation Technology and Automation* (*ICICTA*), Xiangtan, 26-27 October 2019, 166-169.

[13] Cai, Q. () Research on Chinese Naming Recognition Model Based on BERT Embedding. 2019 IEEE 10*th International Conference on Software Engineering and Service Science* (*ICSESS*), Beijing, 18-20 October 2019, 1-4. https://doi.org/10.1109/ICSESS47205.2019.9040736

[14] Gong, C., Tang, J., Zhou, S., *et al*. (2019) Chinese Named Entity Recognition with Bert. 2019 *International Conference on Computer Intelligent Systems and Network Remote Control* (*CISNRC*), Shanghai, 29-30 December 2019, 8-15. https://doi.org/10.12783/dtcse/cisnrc2019/33299

[15] Cui, Y., Che, W., Liu, T., *et al*. (2019) Pre-Training with Whole Word Masking for Chinese Bert. arXiv Preprint arXiv:1906.08101.

[16] Michel, P., Levy, O. and Neubig, G. (2019) Are Sixteen Heads Really Better than One? arXiv:1905.10650.

[17] Wang, Z., Wohlwend, J. and Lei, T. (2019) Structured Pruning of Large Language Models. arXiv Preprint arXiv:1910.04732.

[18] Shen, S., Dong, Z., Ye, J., *et al*. (2019) Q-Bert: Hessian Based Ultra Low Precision Quantization of Bert. arXiv Preprint arXiv:1909.05840.

[19] Radford, A., Narasimhan, K., Salimans, T., *et al*. (2018) Improving Language Understanding by Generative Pre-Training.

[20] Peters, M.E., Neumann, M., Iyyer, M., *et al*. (2018) Deep Contextualized Word Representations. arXiv Preprint arXiv:1802.05365.

[21] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**, 5998-6008.