基于 BERT 的中文命名实体识别方法

王子牛[1] 姜 猛[2] 高建瓴[2] 陈娅先[2]
(贵州大学网络与信息化管理中心 贵阳 550025)[1] (贵州大学大数据与信息工程学院 贵阳 550025)[2]

摘 要 针对传统的机器学习算法对中文实体识别准确率低、高度依赖特征设计以及领域自适应能力差的问题,提出了基于 BERT 的神经网络方法进行命名实体识别。首先,利用大规模未标注语料对 BERT 进行训练,获取文本抽象特征;然后,利用 BiLSTM 神经网络获取序列化文本的上下文抽象特征;最后,通过 CRF 进行序列解码标注,提取出相应的实体。该方法结合 BERT 和 BiLSTM-CRF 模型对中文实体进行识别,以无需添加任何特征的方式在 1998 上半年人民日报数据集上取得了 94.86%的 F1 值。实验表明,该方法提升了实体识别的准确率、召回率及 F1 值,验证了该方法的有效性。

关键词 BERT,命名实体识别,序列标注,BiLSTM,条件随机场
中图法分类号 TP391 **文献标识码** A

Chinese Named Entity Recognition Method Based on BERT

WANG Zi-niu[1] JIANG Meng[2] GAO Jian-ling[2] CHEN Ya-xian[2]
(Network and Information Management Center, Guizhou University, Guiyang 550025, China)[1]
(College of Big Data & Information Engineering, Guizhou University, Guiyang 550025, China)[2]

Abstract In order to solve the problems of low accuracy of traditional machine learning algorithms in Chinese entity recognition, high dependence on feature design and poor adaptability in the field, a recurrent neural network method based on bidirectional encoder representation from transformers was proposed for named entity recognition. Firstly, the BERT is trained by large-scale unlabeled corpus to obtain the abstract features of the text. Then the BiLSTM neural network is used to obtain the contextual features of the serialized text. Finally, the corresponding entities are extracted by sequence labeling with CRF. The method combines the BERT and BiLSTM-CRF models for Chinese entity recognition, and has obtained the F1 value of 94.86% on the People's Daily data set in the first half of 1998 without adding any features. Experiments show that this method improves the accuracy, recall rate and F1 value of entity recognition, indicating the effectiveness of this method.

Keywords BERT, Named entity recognition, Sequence labeling, BiLSTM, Conditional random fields

1 引言

随着大数据时代的快速发展,互联网逐渐成为信息传播的主要方式,网络上每天产生海量的文本数据。为了从海量的数据中提取有用的信息,挖掘其潜在的价值,通常需要用到自然语言处理技术,命名实体识别(Name Entity Recognition, NER)是其中一个重要的组成部分。命名实体识别是信息抽取(Informatica Extraction)中重要的基础部分,其主要任务是识别文本中的人名、地名、组织机构名等专有名词[1],其识别效果对于后续的关系抽取、语义角色标注、自动问答和机器翻译等任务有很大影响。

目前,命名实体识别任务通常被当作序列标注任务,其主要模型分为传统的统计机器学习模型和神经网络模型两类。常见的命名实体识别统计模型主要有隐马尔可夫模型[2](Hidden Markov Model, HMM)和条件随机场(Conditional Random Field, CRF)等[3-5]浅层模型,其中条件随机场模型被广泛应用于各种命名实体识别任务中,并取得了不错的效果。近年来,得益于词向量技术的发展,基于神经网络的深度学习方法在自然语言处理领域中取得了重大的突破。与传统的统计机器学习方法相比,神经网络模型在命名实体识别任务中取得了更好的结果。神经网络方法使用大规模的未标注语料进行词向量训练,通过将预训练的词向量输入到卷积神经网络(Convolutional Neural Network, CNN)、循环神经网络(Recurrent Neural Network, RNN)等模型,实现了端到端的命名实体识别训练。

本文针对中文命名实体识别中传统预训练模型特征提取能力不足且对中文潜在特征表示不充分的问题,将具有更强文本特征表示能力的预训练模型 BERT 作为特征表示层,结合双向长短时记忆网络模型,提取文本全局特征和局部特征,使用人民日报数据集进行实验验证,结果表明本文所提方法有助于提高命名实体识别的整体效果。

2 相关工作

在命名实体识别任务中,早期的信息抽取方法对命名实

本文受贵州省科学技术基金(黔科合 J 字[2015]2045)资助。

王子牛(1961—),男,硕士,副教授,主要研究方向为信息与信号处理、数据挖掘;**姜 猛**(1994—),男,硕士生,主要研究方向为自然语言处理、数据挖掘,E-mail:mjiang_gzu@foxmail.com(通信作者)。

体识别的处理主要是利用先验知识，人工设计出识别模型，然后对模型进行定性和定量的分析和优化[6]。其中，基于规则的方法是通过制定好的规则模板提取相应的信息，该方法需要大量的先验知识，熟悉各实体出现的规律，这将极大地提升任务难度；此外，其还存在时间效率低、可移植性弱等缺点[7]，这种方法在处理结构化单一的数据集上有效，但随着大数据时代的到来，非结构化数据占据很大的比例，对非结构化数据很难获取足够的先验知识来建立特征模板。基于统计机器学习的方法是通过融合语言模型和统计机器学习算法建立模型，如最大熵模型(Maximum Entropy，ME)[8]、隐马尔可夫模型、支持向量机(SVM)[9]和条件随机场等。然而，这些方法的特征提取还是需要人工完成，并且容易失去文本本身的情感信息；在模型训练方面，需要大量人工标注的样本，并且效果也不是特别明显。

基于神经网络的命名实体识别方法通常被当作一个序列标注任务，通过建立序列标注模型对文本进行实体识别。对于序列标注模型，Collobert 等[10]采用 CNN 进行特征抽取，同时提出了一种句子级对数似然函数，通过融合其他特征取得了不错的识别效果。RNN 的提出解决了可变长度输入和如何获取序列前后的长期依赖关系的问题。由 RNN 衍生出来的多种变体，在处理时间序列数据时可以很好地获取和保存序列的上下文信息[11-13]。Huang 等[14]提出了 BiLSTM-CRF 模型，同时还融合了其他语言学特征以提升模型性能。Lample 等[15]采用两个 BiLSTM 分别自动学习词级和字符级特征，在命名实体识别任务中取得了与先进水平可比的结果。Chiu 等[16]提出采用卷积神经网络自动学习字符级特征，在一定程度上缓解了模型对特征工程的依赖，同时还融合了由两个公开的外部资源构造的词典特征，在英文命名实体识别语料上取得了领先的识别结果。Feng 等[17]针对传统的命名实体识别方法构建特征工程和获取相关领域知识的代价昂贵的问题，提出了一种基于 BiLSTM 的神经网络结构的命名实体识别方法。Shen 等[18]提出了利用深度主动学习完成命名实体识别任务的方法，将主动学习与深度学习相结合，可以利用少量的标注数据获得较高的学习准确度。

近年来，深度神经网络在自然语言领域的应用越来越广泛，利用预训练词向量技术，可以避免人工提取特征，直接对原始数据进行处理。在上述命名实体识别方法中，浅层模型采用了特征工程，但在其他领域和语言中的泛化能力不佳；大部分神经网络模型采用了 Word2Vec 工具训练词向量，主要有两种训练方法：Skip-gram 和 CBOW。Skip-gram 是由当前词语预测上下文词语，而 CBOW 是由上下文词语预测当前词语。Word2Vec 尽管能够较好地获得文本序列上下文特征，但仍有很大的挖掘空间。Word2Vec，GloVe 等模型[19]都受限于所使用模型的表征能力，得到的词向量比较偏上下文共现意义，并且未充分考虑词序对词意义的影响。因此，Google 的 Jacob 等[20]提出了基于 Transformer[21]的双向编码器表示(Bidirectional Encoder Representation from Transformers，BERT)方法，它是一个深度双向表示预训练模型，能够更深层次地提取文本的语义信息，在自然语言处理领域具有良好的效果。因此，如何有效地将特征提取能力更强的 BERT 预训练模型和命名实体识别结合起来，从而提升实体的识别性能，成为近期研究的热点。

针对上述问题，本文提出一种基于 BERT 预训练词向量的命名实体识别方法。通过使用 BERT 模型对大规模的文本数据进行训练，利用自注意力机制深度挖掘文本序列之间的潜在特征，从而充分利用文本的上下文信息，进而有助于提升命名实体识别的效果。该方法结合 BiLSTM-CRF 模型，构建了 BERT-BiLSTM-CRF 模型进行中文命名实体识别。该模型是基于 BiLSTM-CRF 模型的改进，将特征提取能力更强的 BERT 模型作为预训练输入，构成了一种完全端到端的、无任何特征工程的命名实体识别模型，在 1998 年人民日报数据集上取得了 94.86%的 F1 值。

3　BERT 模型

当前，在自然语言处理领域，Word2Vec 是使用最广泛的词向量训练工具。Word2Vec 使得深度学习在自然语言处理任务中变得可行，对自然语言处理领域的发展产生了巨大的影响。但 Word2Vec 本身是一种浅层结构价值训练的词向量，所“学习”到的语义信息受制于窗口大小，因此后续有学者提出利用可以获取长距离依赖的 LSTM 语言模型预训练词向量。上述语言模型有自身的缺陷，它是根据句子的上文信息来预测下文，或者根据下文来预测上文。人类理解语言，需要考虑到双向的上下文信息，而传统的 LSTM 模型只学习到了单向的信息。Peters 等[24]提出了 Embeddings from Language Models(ELMo)，ELMo 的出现在一定程度上解决了模型只能学习单向信息的问题。ELMo 是一种双层双向的 LSTM 结构，其训练的语言模型可以学习到句子左右两边的上下文信息。此外，Radford 等[25]提出了 GPT，利用 Transformer 的编码器作为语言模型进行预训练，下游的自然语言处理任务在其基础上进行微调即可。与 LSTM 相比，GPT 语言模型的优点是可以获得句子上下文更远距离的语言信息，但也是单向的。为了充分利用左右两侧的上下文信息，Jacob 等提出了 BERT 模型，图 1 是它与其他模型的结构对比。BERT 模型采用的是双向 Transformer，它的特征表示在所有层中共同依赖于左右两侧的上下文，该模型融合了其他模型的优点，并摒弃了它们的缺点，在诸多自然语言处理的后续特定任务上取得了良好的效果。

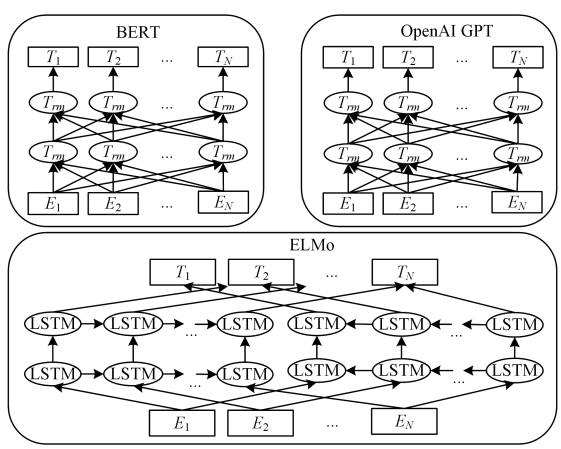

图 1　预训练模型结构对比

BERT 是一种基于微调的多层双向 Transformer 编码器，由于该模型需要海量的数据以及强大的计算能力才能实现训练，因此谷歌开源实现了两个版本的 BERT 模型。在两个版本中，前馈大小都设置为 4 层，其参数如表 1 所列。

表 1 模型参数

	L	H	A	Total Parameters/M
BERT Base	12	768	12	110
BERT Large	24	1024	16	340

其中，L 表示层数(即 Transformer blocks 块)，隐藏大小表示为 H，A 是自注意力的“头数”，Total Parameters 是模型的所有参数。

BERT 模型的两个版本的本质是一样的；区别在于参数的设置。BERT Base 作为 baseline 模型，在此基础上优化，进而出现了 BERT Large。本文实验使用的是谷歌针对中文语料训练好的 BERT Base 版本。

3.1 输入表示

输入表示可以在一个词序列中表示单个文本句或一对文本，例如：[问题，答案]。对于给定的词，其输入表示可以通过 3 部分 Embedding 求和组成。Embedding 的可视化表示如图 2 所示。

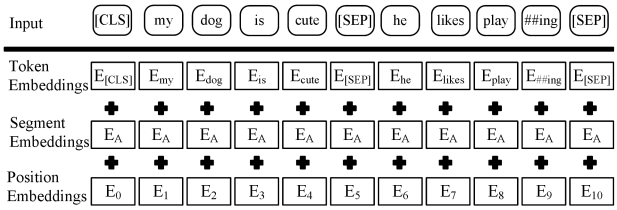

图 2 BERT 模型的输入表示

其中，Token Embeddings 表示的是词向量，在中文处理中可以是词向量或字向量，本文实验所用的是更符合中文特征的字向量；第一个单词是 CLS 标志，可以用于之后的分类任务，对于非分类任务，可以忽略词向量；做以两个句子为输入的分类任务时，用 Segment Embeddings 来区别两种句子；Position Embeddings 是通过模型学习得到的位置信息。

3.2 BERT 模型预训练任务

BERT 模型使用两个新的无监督预测任务对 BERT 进行预处理，分别是 Masked Language Model(Masked LM)和下一句预测。

3.2.1 Masked LM

预训练的目标是构建语言模型，BERT 模型采用的是双向 Transformer。为了训练深度双向 Transformer 表示，采用一种简单的方法：随机掩盖部分输入词，然后对被掩盖的词进行预测。

在训练的过程中，随机地掩盖每个序列中 15% 的标签，与 Word2Vec 中的 CBOW 对每一个词都进行预测不同，Masked LM 从输入中随机地掩盖一些词，其目标是基于上下文来预测被掩盖单词的原始词汇。与从左到右的语言模型预训练不同，Masked LM 学习到的表示能够融合左右两侧的上下文。模型中的双向 Transformer 不知道它将被要求预测哪些词，或者哪些已经被随机词替换，因此它必须对每个输入词保持分布式的上下文表示。此外，随机替换在所有词中只发生 1.5%，并不会影响模型对语言的理解。

3.2.2 下一句预测

自然语言中的很多句子级别的任务，如自动问答(Question Answering，QA)和自然语言推理(Natural Language Inference，NLI)等任务，都需要理解两个句子之间的关系。例如，上述 Masked LM 任务中，经过第一步的处理，1.5% 的词汇被遮盖，那么在这一任务中，需要随机将数据划分为同等大小的两部分，一部分数据中的两个语句对是上下文连续的，另一部分数据中的两个语句对是上下文不连续的，然后让 Transformer 模型来识别这些语句对，判断下一句与当前句是否连续。下一句预测的格式如表 2 所列。

表 2 下一句预测格式

输入	标签
[CLS] 这个男人去了[MASK] 商店 [SEP] 他 买 了 一 加 仑[MASK] 牛 奶 [SEP]	IsNext
[CLS] 这个男人[MASK] 这 家 商店 [SEP] 企 鹅 [MASK] 是飞行## 小鸟[SEP]	NotNext

4 基于 BiLSTM-CRF 的命名实体识别模型

4.1 BiLSTM 结构

给定输入序列 $(x_1,x_2,\cdots,x_n)$，循环神经网络模型返回一个关于输入序列的表示序列 $(h_1,h_2,\cdots,h_n)$。RNN 模型能动态地捕获序列数据的信息，并且对信息有记忆保存的能力，但在算法实现上容易产生梯度消失或爆炸问题。Hochreiter 等[22]提出的 LSTM 模型引入了记忆单元和门限机制，实现了对长距离信息的有效利用，并解决了梯度消失问题。Graves 等[23]对 LSTM 模型的记忆单元和门限机制做了改进以提高效率，本文采用 Graves 等提出的改进门限机制，在 t 时刻，给定输入 x_t，LSTM 的隐藏层输出表示的具体计算过程如式(1)－式(5)所示：

$$i_t=\sigma(\boldsymbol{W}_{xi}x_t+\boldsymbol{W}_{hi}h_{t-1}+\boldsymbol{W}_{ci}C_{t-1}+\boldsymbol{b}_i) \tag{1}$$

$$f_t=\sigma(\boldsymbol{W}_{xf}x_t+\boldsymbol{W}_{hf}h_{t-1}+\boldsymbol{W}_{cf}C_{t-1}+\boldsymbol{b}_f) \tag{2}$$

$$C_t=f_tC_{t-1}+i_t\tanh(\boldsymbol{W}_xCx_t+\boldsymbol{W}_{hC}h_{t-1}+\boldsymbol{b}_C) \tag{3}$$

$$o_t=\sigma(\boldsymbol{W}_{xo}x_t+\boldsymbol{W}_{ho}h_{t-1}+\boldsymbol{W}_{co}C_t+\boldsymbol{b}_o) \tag{4}$$

$$h_t=o_t\tanh(C_t) \tag{5}$$

其中，$\boldsymbol{W}$ 表示连接两层的权重矩阵(如 $\boldsymbol{W}_{xi}$ 表示输入层到隐藏层的输入门的权重矩阵)，$\boldsymbol{b}$ 表示偏置向量(如 $\boldsymbol{b}_i$ 表示隐藏层的输入门的偏置向量)，C 表示记忆单元的状态，σ 和 tanh 表示两种不同的神经元激活函数，i_t，f_t 和 o_t 分别表示输入门、遗忘门和输出门。这种门限机制能够对记忆单元的信息进行有效过滤和记忆，从而解决 RNN 存在的问题。

4.2 CRF 模型

通常，在预测阶段采用 softmax 分类器解决多分类问题，但 softmax 分类器在序列标注问题中没有考虑到标签之间的依存关系。因此，本文方法使用 Collobert 等提出的句级对数似然函数，也就是 CRF 模型。该方法能考虑标签序列的全局信息，更好地对标签进行预测。具体细节如下：

假定引入转移得分矩阵 $\boldsymbol{A}$，矩阵元素 $\boldsymbol{A}_{i,j}$ 表示标签 i 转移到标签 j 的转移得分，令 y_0 和 y_{n+1} 为句中的起始标签和终止标签，标签种类为 k，则 $\boldsymbol{A}\in\boldsymbol{R}^{(k+2)*(k+2)}$。设句子长度为 n，则输出层的得分矩阵为 $\boldsymbol{P}\in\boldsymbol{R}^{n*k}$，矩阵元素 $P_{i,j}$ 表示第 i 个词在第 j 个标签下的输出得分；给定输入句子 $X=(x_1,x_2,\cdots,x_n)$，输出标签序列 $y=(y_1,y_2,\cdots,y_n)$，则该标签序列的总得分为：

$$S(X,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{6}$$

对所有可能的序列路径进行归一化，产生关于输出序列 y 的概率分布，如式(7)所示：

$$P(y|X)=\frac{e^{S(X,y)}}{\sum_{\tilde{y}\in Y_X} e^{S(X,\tilde{y})}} \tag{7}$$

在训练过程中，最大化关于正确标签序列 y^* 的对数概率，如式(8)所示：

$$\log(P(y^*|X))=S(X,y^*)-\log(\sum_{\tilde{y}\in Y_X} e^{S(X,\tilde{y})}) \tag{8}$$

由式(8)可知，采用句级似然函数的目的是鼓励模型生成正确的标签序列。在解码阶段，预测总得分最高的序列作为最优序列，如式(9)所示：

$$y^*=\arg\max_{\tilde{y}\in Y_X} S(X,\tilde{y}) \tag{9}$$

在预测阶段，采用动态规划算法 Viterbi 来求解最优序列。

4.3 BERT-BiLSTM-CRF 命名实体识别模型

BERT-BiLSTM-CRF 模型是将 BiLSTM 网络和 CRF 模型结合起来，即在 BiLSTM 网络的隐藏层后加一层 CRF 线性层，模型结构如图 3 所示。

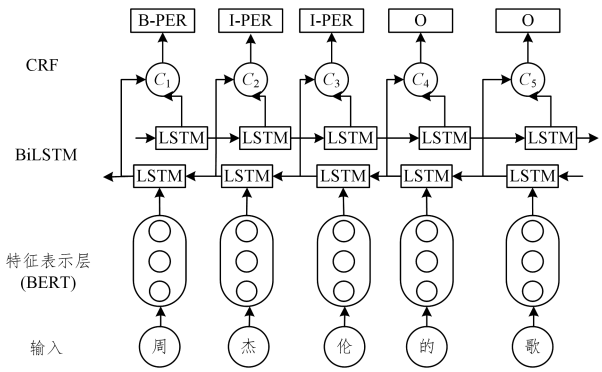

图 3　BERT-BiLSTM-CRF 模型结构

双向 LSTM 的基本思想就是对每一个训练序列分别作用一个向前和向后的 LSTM 网络，并且这两个连接着同一个输出层，这样的一种网络结构可以给输出层提供每一个序列点完整的上下文信息，并且经过 CRF 有效地考虑了序列前后的标签信息。

由图 3 可以看出，BERT-BiLSTM-CRF 命名实体识别模型是通过使用 BERT 模型作为特征表示层加入到双向 LSTM 模型中。BiLSTM 获取句子表示的过程可以用算法 1 描述。

算法 1　BiLSTM 获取句子表示

输入：分词结果、词向量、前向 LSTM 输出的前文信息向量、后向 LSTM 输出的后文信息向量

输出：句子表示向量 $\mathbf{o}_t$

Step1　将分词后的序列输入到 BERT 特征表示层，得到句子的向量表示。

Step2　对于句子中的每个词语 W_i，将其词向量 $\mathbf{e}_i(W_i)$ 输入到 BiLSTM 网络中，计算当前状态隐藏层的值。

更新输入门：$i_t=\sigma(\mathbf{W}_{xi}x_t+\mathbf{W}_{hi}h_{t-1}+\mathbf{W}_{ci}C_{t-1}+\mathbf{b}_i)$

更新遗忘门：$f_t=\sigma(\mathbf{W}_{xf}x_t+\mathbf{W}_{hf}h_{t-1}+\mathbf{W}_{cf}C_{t-1}+\mathbf{b}_f)$

更新输出门：$o_t=\sigma(\mathbf{W}_{xo}x_t+\mathbf{W}_{ho}h_{t-1}+\mathbf{W}_{co}C_t+\mathbf{b}_o)$

计算记忆单元的值：$C_t=f_tC_{t-1}+i_t\tanh(\mathbf{W}_{xC}x_t+\mathbf{W}_{hC}h_{t-1}+\mathbf{b}_C)$

计算 t 时刻隐藏层的值：$h_t=o_t\tanh(C_t)$。

Step3　将前向信息向量与后向信息向量按位拼接，得到句子表示 $h_t=[\overrightarrow{h_t};\overleftarrow{h_t}]$。

5　实验与分析

5.1　实验数据及评价指标

为验证所提模型的有效性，本文采用北京大学计算语言学研究所发布的 1998 年上半年的人民日报语料进行相关实验研究。1998 年人民日报语料已经分好词，标注了人名、地名、机构名等信息。本文使用三元标记集{B,I,O}，B 表示实体的第一个词，I 表示机构名的其余词，O 表示不属于机构名的词。人名记为 PER，所以人名的开始记为 B-PER；地名记为 LOC，地名的开始记为 B-LOC；机构名记为 ORG，机构名的开始记为 B-ORG。在实验中，以 1998 年 6 月份的数据为测试集，1998 年 1 月－5 月的数据为训练集。其中，训练集与测试集的信息统计如表 3 所列。

表 3　1998 年人民日报预料数据统计

	字数	LOC	PER	ORG
训练集	31944544	12471	22665	52118
测试集	660970	2770	4408	11228

对每一类命名实体，都采用准确率(Precision，P)、召回率(Recall，R)以及调和平均数 F1 值(F1-score)作为模型性能的评价标准。定义如下：

$$P=\frac{\text{正确识别出的命名实体个数}}{\text{识别出的命名实体个数}}\times 100\% \tag{10}$$

$$R=\frac{\text{正确识别出的命名实体个数}}{\text{标准结果中命名实体个数}}\times 100\% \tag{11}$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\% \tag{12}$$

5.2　模型搭建和参数设置

本文所提模型采用 Tensorflow 进行搭建。Tensorflow 是由谷歌人工智能团队开发的深度学习框架，被广泛应用于各类机器学习算法的实现。实验参数设置如下：输入维度 *seq_length* 为 128，训练集的 *batch_size* 为 64，测试集的 *batch_size* 为 8，训练学习率为 2×10^{-5}。为了防止训练中出现梯度爆炸，使用梯度夹子(Gradient Clipping)技术并设置参数为 5，使用 dropout 技术来防止过拟合，值设为 0.5。

5.3　实验结果

在数据集上，采用 CRF，CNN，LSTM，BiLSTM，BiLSTM-CRF 和 BERT-BiLSTM-CRF 模型进行性能分析。实验结果如表 4 所列。

表 4　模型在 1998 年人民日报数据集上的识别效果对比

(单位：%)

任务	准确率	召回率	F1 值
CRF	85.17	85.56	85.36
CNN	86.92	85.16	86.03
LSTM	87.25	85.76	86.49
BiLSTM	88.34	87.61	87.97
BiLSTM-CRF	90.45	89.72	90.08
BERT-BiLSTM-CRF	94.73	94.99	94.86

从表 4 中可以看出，在 CRF 模型和其他基于神经网络的模型比较中，基于神经网络的模型在各个方面的性能都好于 CRF 模型；在 CNN 模型与其他基于 RNN 实现的模型比较中，基于 RNN 的方法整体上优于 CNN，CNN 在图像处理领域表现良好，对于文本序列，基于 RNN 模型有更好的表现；从 LSTM 模型和 BiLSTM 模型的对比中可以看出，BiLSTM 的表现优于 LSTM，验证了双向的 LSTM 模型对序列的上下

文信息的学习能力更强;在 BiLSTM 和 BiLSTM-CRF 模型的比较中,利用 CRF 来进行序列标注的 BiLSTM-CRF 模型在各方面的表现都优于 BiLSTM,说明 CRF 在解码过程中考虑了序列中的全局标签信息,因此提升了模型性能。在 BiLSTM-CRF 模型和 BERT-BiLSTM-CRF 的比较中,后者相比前者有了不低于 5%的性能提升,可以看出 BERT 模型充分刻画了文本数据中字符间的关系特征,有助于提升模型性能。

为了验证本文方法的有效性,将其与其他主流命名实体识别方法进行实验对比,具体结果如表 5 所列。

表 5 本文方法与主流方法的识别效果对比

(单位:%)

方法	准确率	召回率	F1 值
Collobert 等[10]	88.43	87.68	88.05
Lample 等[15]	90.45	89.72	90.08
Chiu 等[16]	91.94	90.06	90.99
Shen 等[18]	91.46	90.18	90.81
本文方法	94.73	94.99	94.86

表 5 为本模型与其他公开方法在 1998 年人民日报语料上的比较。Collobert 等采用前馈神经网络,结合预处理和词缀、大小写特征,取得了 F1 值为 88.05%的结果;Lample 等将字符级词向量输入双向 LSTM-CRF 模型,达到了 90.08%的 F1 值;Chiu 等将 BiLSTM 与 CNN 模型结合,取得了 91.49%的先进结果。本模型同为 BiLSTM-CRF 模型,仅改变输入的预训练词向量模型,获得了 94.86%的 F1 值,优于 Lample 和 Chiu 等人的结果。实验结果表明,相比于 BiLSTM 和 CNN,基于 Transformer 的 BERT 模型具有更强的特征提取能力,这与 Transformer 中的自注意力机制选择信息特征的能力有着密不可分的关系,使得本文模型的性能有更好的表现。

结束语 本文构造了一种完全端到端、无需人工特征的神经网络模型 BERT-BiLSTM-CRF,并将其用于命名实体识别任务。该模型将使用海量的文本数据和深层网络模型训练出来的预训练词向量作为输入,使模型充分学习文本的特征信息,同时增强了字符间的推理能力,提升了实体识别的效果。该模型既结合了 BiLSTM 学习词语的上下文信息的能力,同时又保留了 CRF 模型通过全局信息推断标签的能力,在 1998 年人民日报语料上取得了 94.86%的结果,超过了当前大部分公开的方法,达到了当前较为先进的水平。实验表明,基于 BERT 的命名实体识别模型能够提升实体识别的结果,对后续的研究具有一定的参考价值。下一步将考虑引入语言学特征,结合该模型进行联合训练,以提升模型性能。

参 考 文 献

[1] SANG,ERIK F,KIM T,et al. Introduction to the CoNLL-2003 Shared Task:Language-Independent Named Entity Recognition [J]. IEEE TransactionsonWireless Communications, 2003, 21(8):142-147.

[2] BENGIO Y,SCHWENK H,SENÉCAL J S. Neural Probabilistic Language Models[J]. Journal of Machine Learning Research, 2001,3(6):1137-1155.

[3] 赵晓凡,赵丹,刘永革. 利用 CRF 实现中文人名性别的自动识别[J]. 微电子学与计算机,2011,28(10):122-128.

[4] 林广和,张绍武,林鸿飞. 基于细粒度词表示的命名实体识别研究[J]. 中文信息学报,2018,32(11):62-71.

[5] LUO G,HUANG X J,LIN C Y,et al. Joint named entity recognition and disambiguation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. 2015:1030-1038.

[6] ZHANG L,ZHANG Y. Big Data Analysis by Infinite Deep Neural Networks [J]. Journal of Computer Research and Development,2016,53(1):68-79.

[7] ZHANG S,WANG X. Automatic Recognition of Chinese Organization Name Based on Conditional Random Fields[C]// International Conference on Natural Language Processing and Knowledge Engineering,2007. IEEE,2007:229-233.

[8] BORTHWICK A E. A maximum entropy approach to named entity recognition[M]. New York:New York University,1999.

[9] 程健一,关毅,何彬. 基于 SVM 和 CRF 双层分类器的英文电子病历去隐私化[J]. 智能计算机与应用,2016,6(6):17-19.

[10] COLLOBERT R,WESTON J,BOTTOU L,et al. Natural language processing (almost) fromscratch[J]. Journal of Machine Learning Research,2011,12(1):2493-2537.

[11] GRAVES A. Long Short-Term Memory[M]// Supervised Sequence Labelling with Recurrent Neural Networks. Springer Berlin Heidelberg,2012:1735-1780.

[12] 殷昊,徐健. 基于字词融合特征的微博情绪识别方法[J]. 计算机科学,2018,45(S2):105-109.

[13] CHO K,VAN MERRIENBOER B,GULCEHRE C, et al. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation[J]. arXiv: 1406. 1078v3, 2014.

[14] HUANG Z,XU W,YU K. BidirectionalLSTM-CRF models for sequencetagging[J]. arXiv:1508.01991,2015.

[15] LAMPLE G,BALLESTEROS M,SUBRAMAN S,et al. Neural architectures for namedentity recognition[C]// Proceedings of NAACL-HLT. 2016:260-270.

[16] CHIU J P,NICHOLS E. Namedentity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics,2015,4(10):357-370.

[17] 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.

[18] SHEN Y,YUN H,LIPTON Z C,et al. Deep Active Learning for Named Entity Recognition[J]. arXiv:1707.05928v3,2018.

[19] JEFFREY P,SOCHER R,MANNING C D. GloVe:Global Vectors for Word Representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014:1532-1543.

[20] DEVLIN J,CHANG M W,LEE K,et al. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. arXiv:1810.04805v1,2018.

[21] VASWANI A,NOAM S,et al. Attention Is All You Need[J]. arXiv:1706.03762v5,2017.

[22] HOCHREITER S,SCHMIDHUBER J. Long Short-term Memory[J]. Neural Computation,1997,9(8):1735-1780.

[23] GRAVES A,MOHAMED A R,HINTON G. Speech Recognition with Deep Recurrent Neural Networks[C]// IEEE International Conference on Acoustics,Speech and Signal Processing. 2013:6645-6649.

[24] PETERS M E,NEUMANN M. Deep contextualized word representations[J]. arXiv:1802.05365v2,2018.

[25] RADFORD A,NARASIMHAN K. Improving Language Understanding by Generative Pre-Training[J/OL]. https://s3-us-west2. amazonaws. com/openai-assets/research-covers/language-unsupervised/language_unders-tanding_paper. pdf.