计算机科学
COMPUTER SCIENCE

基于MacBERT和对抗训练的审计文本命名实体识别

钱泰羽, 陈一飞, 庞博文

引用本文

钱泰羽, 陈一飞, 庞博文. 基于MacBERT和对抗训练的审计文本命名实体识别[J]. 计算机科学, 2023, 50(11A): 230200083-6.

QIAN Taiyu, CHEN Yifei, PANG Bowen. Audit Text Named Entity Recognition Based on MacBERT and Adversarial Training [J]. Computer Science, 2023, 50(11A): 230200083-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230200083

基于 MacBERT 和对抗训练的审计文本命名实体识别

钱泰羽　陈一飞　庞博文
南京审计大学计算机学院　南京 211815
(qiantyy@163.com)

摘　要　为了从审计文本中自动识别有效的实体信息,提高政策跟踪审计的效率,提出一种基于 MacBERT(MLM as correction BERT)和对抗训练的审计文本命名实体识别(Named Entity Recognition,NER)模型(Audit-MBCA)。目前深度学习在 NER 任务上应用成熟且成果显著,但审计文本存在语料库缺乏、实体边界识别不清晰等问题。针对这些问题,文中构建了审计文本数据集并将其命名为 Audit 2022,使用 MacBERT 中文预训练语言模型获得其向量表示,同时引入对抗训练,利用中文分词(Chinese Word Segmentation,CWS)任务与 NER 任务的共享词边界信息帮助进行实体边界识别。实验结果表明,Audit-MBCA 模型在 Audit 2022 数据集上的 $F1$ 值为 91.05%,较主流模型提升了 4.53%;在 SIGHAN 2006 数据集上的 $F1$ 值为 93.70%,较其他模型提升了 0.33%～3.25%,验证了所提模型的有效性和泛化能力。

关键词:审计文本;命名实体识别;MacBERT;对抗训练
中图法分类号　TP391

Audit Text Named Entity Recognition Based on MacBERT and Adversarial Training

QIAN Taiyu,CHEN Yifei and PANG Bowen
School of Computer Science,Nanjing Audit University,Nanjing 211815,China

Abstract　In order to automatically identify the effective entity information from the audit text and improve the efficiency of policy tracking audit,a named entity recognition(NER) of audit text model(Audit-MBCA) based on MacBERT(MLM as correction BERT) and adversarial training is proposed. At present,deep learning has been maturely applied to NER task and achieved significant results. However,the audit text has some problems such as lacking corpus and unclear entity boundary recognition. To address these problems,the audit text dataset named Audit2022 is constructed in this paper. Its vector representation is obtained by using the MacBERT Chinese pre-training language model. At the same time,adversarial training is introduced and the shared word boundary information of Chinese word segmentation(CWS) task and NER task is used to help identify entity boundaries. Experimental results show that the value of F1 on the Audit2022 dataset from the Audit-MBCA model is 91.05%,which is 4.53% higher than the mainstream model;the value of F1 on the SIGHAN2006 dataset is 93.70%,which is 0.33%～3.25% higher than other models. These verify the effectiveness and generalization ability of the proposed model.

Keywords　Audit text,Named entity recognition,MacBERT,Adversarial training

1　引言

随着新审计法的颁布,审计政策划分越来越详尽,重要政策文件、项目会议纪要、官方新闻报道等反映政策传达情况的审计文本[1]也在逐日增加,同时,国家重大政策措施落实情况跟踪审计(以下简称“政策跟踪审计”)的监督作用也日益重要。但现有的政策跟踪审计自动化程度不完善,多以人工为主,且在政策跟踪审计过程中政策执行会不断产生新的审计文本,这不仅加大了审计人员的工作量、加剧了审计过程中的时间消耗,且容易忽略审计文本内容之间的相关性。命名实体识别(NER)是自然语言处理(Natural Language Processing,NLP)最重要的基础任务,能够自动从审计文本中识别出与审计相关的人名、地名、机构名和专有名词等具有特定意义的实体类型,并利用它们进行关系提取、构建知识图谱等,从而帮助政策跟踪审计在事前、事中和事后审计时,能够更好地掌握审计文本之间的关联信息。

早期的基于规则和词典[2]的命名实体识别方法,需要构造特定的规则模板,尤其依赖于知识库和词典的建立,存在可移植性差等问题。基于统计机器学习的命名实体识别方法,需要依赖对特征的选取和分析,并将影响任务因素较大的特征加入特征模板中,主要包括隐马尔可夫模型(HMM)[3]、最大熵模型(MEM)[4]、支持向量机(SVM)[5]和条件随机场(CRF)[6]等。随着深度学习的出现,大量的模型被应用到命名实体识别中并取得了优异的成绩。相较于基于统计机器学习的方法,其能够有效避免复杂的人工特征抽取,具有良好的泛化能力。Hammerton[7]最先将长短期记忆网络(Long Short-Term Memory,LSTM)与 CRF 相结合,并将其应用到命名实体识别上,该网络的序列建模能力优良,因此 LSTM-CRF 成为了实体识别的基础模型。在该模型的基础上,Lample 等[8]提出双向长短期记忆网络(Bi-directional Long Short-

基金项目:江苏省研究生科研与实践创新计划项目(SJCX22_0995)
This work was supported by the Postgraduate Research & Practice Innovation Program of Jiangsu Province(SJCX22_0995).
通信作者:陈一飞(yifeichen91@nau.edu.cn)

Term Memory, BiLSTM)与 CRF 相结合的网络模型，这种网络模型能够获得上下文的双向序列信息，因此被广泛应用于命名实体识别任务中。Chi[9]将领域知识表示融合到 BiLSTM-CRF 模型中进行实体识别，并用于构建问答系统，模型在企业财务审计领域语料上的 $F1$ 值为 83.56%。随着深度学习的不断发展，模型参数显著增加，需要越来越大的数据集用于充分训练模型参数并预防过拟合。预训练语言模型[10]可以在海量文本中通过预训练学习到一种通用语言表示，提供更好的模型初始化，从而具有更好的泛化性并有助于完成下游任务。

尽管这些深度学习方法在命名实体识别上已取得很好的成绩，但在审计文本命名实体识别过程中仍存在以下问题。首先，目前针对审计领域识别研究的还很少，且缺乏相应的语料库。其次，现有的预训练语言模型在各任务上性能表现优异，但却多以英文为基础且没有根据中文语言特点进行优化。最后，由于中文词不具备天然的界限，实体边界识别不清晰会导致模型性能下降，在对审计文本进行实体识别时需要确定实体边界。中文分词(CWS)任务用于识别词边界，具有丰富的训练数据量，且与 NER 任务有许多相同的词边界信息，有助于 NER 任务进行实体边界识别。如表 1 所示，给定句子“咸丰县成立精准扶贫作战指挥部”，其中“咸丰县”和“成立”的词边界在 NER 任务和 CWS 任务中相同，为任务的共享信息；而对于“精准扶贫作战指挥部”，NER 任务比 CWS 任务的边界粒度更粗，则为任务的私有信息。

表 1 NER 任务和 CWS 任务的对比

Table 1 Comparison between NER and CWS tasks

任务	咸丰县成立精准扶贫作战指挥部
NER	咸丰县|成立|精准扶贫作战指挥部 B-LOC I-LOC I-LOC|O O|B-ORG I-ORG I-ORG I-ORG I-ORG I-ORG I-ORG I-ORG I-ORG
CWS	咸丰县|成立|精准|扶贫|作战|指挥部 B I E|B E|B E|B E|B E|B I E

针对上述问题，本文提出基于中文预训练语言模型和对抗训练的审计文本命名实体识别方法。其主要贡献如下：

(1)构建了审计文本数据集并命名为 Audit 2022。首先使用网络爬虫技术获取审计文本数据，其次对数据进行预处理和划分，最后对数据进行标注和校验。目前尚未有审计文本命名实体识别语料库。

(2)更好的实体边界识别模型。使用中文预训练语言模型获得审计文本的向量表示，并将对抗训练应用到审计文本命名实体识别任务中。本文模型在 Audit 2022 数据集上的 $F1$ 值为 91.05%，较主流模型提升了 4.53%；在 SIGHAN 2006 数据集上的 $F1$ 值为 93.70%，较其他模型提升了 0.33%～3.25%，验证了本文模型的有效性和泛化能力。

2 相关工作

2.1 预训练语言模型

BERT(Bidirectional Encoder Representation from Transformers)是于 2018 年 10 月由 Google AI 研究院的 Devlin 等[11]提出的一种预训练语言模型。该模型采用 Transformer 进行编码，引入了自注意力机制(Self-Attention)预测词间的依赖关系及捕获句子内部结构的信息，刷新了当时 NLP 任务中的诸多最佳记录。同时，BERT 的提出也解决了 Word2vec 无法区分一词多义的问题[12]。尽管 BERT 在各任务上表现优异，但却以英文为预训练基础，在随机掩码过程中没考虑中文词级特征。Cui 等[13]提出了中文预训练语言模型 MacBERT(MLM as correction BERT)，并在相关中文 NLP 数据集上证明了模型的性能。MacBERT 在 BERT 的基础上进行了改进：一是引入了纠错型掩码语言模型(MLM as correction, Mac)预训练任务，使用全词掩码策略(Whole Word Masking, WWM)对该字所在的词进行全部掩码，同时使用 N-gram 策略来决定被[MASK]标记的比例，如果 N-gram 被选中进行掩码，则会查找相似的词进行掩码，若没有相似的词，则降级使用随机词进行替换，从而缓解“预训练-下游任务”不一致的问题；二是采用句子顺序预测(Sentence Order Prediction, SOP)任务替换 NSP(Next Sentence Prediction)任务，SOP 任务是使模型判断出两个句子的正确顺序。因此 MacBERT 可以更好地适用于获得中文数据集的向量表示。Jiao 等[14]将 MacBERT 应用到反恐领域细粒度实体识别任务中，并通过实验证明了该中文预训练语言模型的有效性。文献检索发现目前还没有研究将 MacBERT 应用到审计文本中。

2.2 对抗训练

生成对抗网络(Generative Adversarial Networks, GAN)由 Goodfellow 等[15]最早提出并用于图像领域。GAN 由任务鉴别器和任务生成器构成，任务鉴别器用于判断一个样本是真实样本还是任务生成器的生成样本，而任务生成器则尽量生成任务鉴别器无法判断是否是其生成的样本。近期的研究将对抗训练应用到 NLP 领域中，并取得了很好的成绩。Cao 等[16]提出一种对抗迁移学习框架并结合自注意力机制，将 CWS 任务中的共享词边界信息整合到 NER 任务中，同时防止 CWS 任务的私有信息带来的噪声，在 SIGHAN 2006 数据集上的 $F1$ 值为 90.64%。Zhang[17]使用 DeepCAN 作为共享-私有特征提取器进行特征提取，同时使用选择卷积注意力网络作为生成对抗网络的鉴别器，与共享特征提取器 DeepCAN 进行对抗学习，提取两个任务共享的词边界信息，在 SIGHAN 2006 数据集上的 $F1$ 值为 91.82%。尽管他们在各自的模型上取得了不错的结果，但均没能结合预训练语言模型。

3 基于 MacBERT 和对抗训练的命名实体识别模型

为了获得审计文本更好的向量表示和解决实体边界识别不清晰的问题，本文提出一种基于 MacBERT 和对抗训练的审计文本命名实体识别模型(Audit-MBCA)。如图 1 所示，该模型包括 3 个任务，分别是命名实体识别(NER)任务、中文分词(CWS)任务和对抗训练任务。

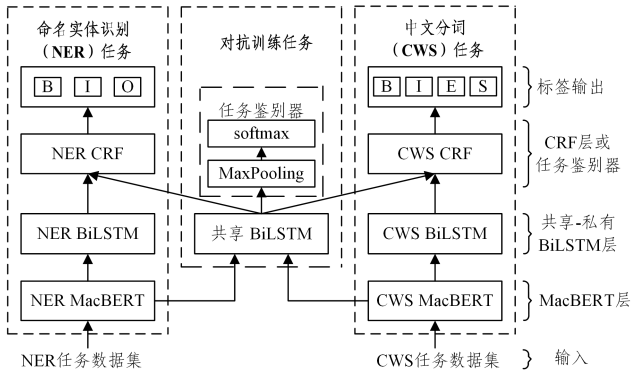

图 1 审计文本命名实体识别框架图

Fig. 1 Framework diagram of audit text named entity recognition

3.1 NER 任务和 CWS 任务

NER 任务和 CWS 任务纵向均包括 MacBERT，BiLSTM 和 CRF 模块。由于 NER 任务为主任务，CWS 任务为辅助任务，下面将以 NER 任务为主且横向对模型结构进行介绍。

3.1.1 MacBERT 层

本文使用 MacBERT 获得输入文本的向量表示。MacBERT 由 Word 嵌入、Positional 嵌入、Token type 嵌入和一个连续的 L 层 Transformer 构成。Word 嵌入用于表示字本身的信息，Positional 嵌入用于编码和学习字在句子中的位置信息，Token type 嵌入用于判断给定句子间是否是连续的方式获得句子级别特征，最后将 Word 嵌入、Positional 嵌入和 Token type 嵌入获得的向量相加。给定输入句子 $C=\{c_1,c_2,\cdots,c_n\}$，在句子首位和末位分别添加[CLS]标签和[SEP]标签，经过 MacBERT 处理后，得到输入句子 C 的向量表示为 $\{\boldsymbol{x}_0,\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{n+1}\}$。

3.1.2 BiLSTM 层

长短期记忆网络是循环神经网络(Recurrent Neural Network，RNN)的一种变体，它可以有效利用长距离信息，通过门控结构和记忆单元解决 RNN 在训练过程中出现的梯度弥散或梯度爆炸的问题。LSTM 单元结构如图 2 所示。

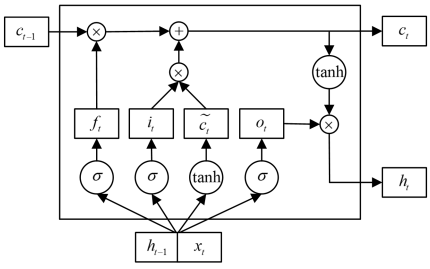

图 2 LSTM 单元结构

Fig. 2 LSTM cell structure

LSTM 单元结构由 3 个门控单元和 1 个记忆单元组成，其中 3 个门控单元包括遗忘门、输入门和输出门。LSTM 能够记忆长期依赖的关键在于遗忘门与输入门，遗忘门能够决定遗弃什么样的信息，输入门能够决定保留什么样的信息，输出门能够决定输出多少信息。计算式如式(1)－式(3)所示：

$$\begin{bmatrix} f_t \\ i_t \\ o_t \\ \tilde{c}_t \end{bmatrix}=\begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix}\left(\boldsymbol{W}\begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}+\boldsymbol{b}\right) \tag{1}$$

$$c_t=i_t\times\tilde{c}_t+f_t\times c_{t-1} \tag{2}$$

$$h_t=o_t\times\tanh(c_t) \tag{3}$$

其中，f_t，i_t，o_t 分别为第 t 时刻的遗忘门、输入门和输出门，$\tilde{c}_t$ 为 t 时刻的状态，c_t 为 t 时刻记忆单元的状态，h_t 为 t 时刻的输出，σ 为 sigmoid 激活函数，tanh 为双曲正切激活函数，$\boldsymbol{W}$ 为权重，$\boldsymbol{b}$ 为偏置项。

由于 LSTM 只能获得当前时刻输入信息的前一刻信息，在序列标注任务中，当前时刻输入信息的后一刻信息同样至关重要。为了融合序列两侧的信息，本文采用 BiLSTM 进行特征提取。计算式如式(4)－式(6)所示：

$$\overrightarrow{h_t}=\overrightarrow{LSTM}(\overrightarrow{h_{t-1}},x_t) \tag{4}$$

$$\overleftarrow{h_t}=\overleftarrow{LSTM}(\overleftarrow{h_{t-1}},x_t) \tag{5}$$

$$h_t=\overrightarrow{h_t}\oplus\overleftarrow{h_t} \tag{6}$$

其中，$\overrightarrow{h_t}$ 和 $\overleftarrow{h_t}$ 分别为第 t 时刻的前向和后向的隐藏状态，$\oplus$ 表示连接操作。

3.1.3 CRF 层

BiLSTM 仅可获得上下文的信息关系，但不会考虑连续标签之间的相互关系，因此本文将 BiLSTM 训练输出的特征使用 CRF 进行标签序列预测，但 BiLSTM 输出向量的维度与 CRF 之间不对等，为了便于 CRF 进行标签序列预测时对损失函数进行计算，在 BiLSTM 层后添加一个全连接层。计算式如式(7)所示：

$$p_i=\boldsymbol{W}_p h_i+\boldsymbol{b}_p \tag{7}$$

其中，$\boldsymbol{W}_p$ 为权重，$\boldsymbol{b}_p$ 为偏置项。

由于 NER 任务和 CWS 任务的标签输出不同，因此为两个任务分配各自的 CRF 层，从而得到各自任务的标签序列。CRF 预测标签序列的计算式如式(8)所示：

$$Score(X,Y)=\sum_{i=1}^{n}O_{i,y_i}+\sum_{i=1}^{n}P_{y_{i-1},y_i} \tag{8}$$

其中，X 为输入序列，Y 为预测标签序列，O_{i,y_i} 为字符 x_i 被标记为第 y_i 个标签的得分，P_{y_{i-1},y_i} 为 y_{i-1} 标签转移 y_i 标签的得分，n 为句子的长度。

预测标签序列 Y 产生的概率，计算式如式(9)所示：

$$P(Y\mid X)=\frac{e^{Score(X,Y)}}{\sum\limits_{\tilde{Y}\in Y_X}e^{Score(X,\tilde{Y})}} \tag{9}$$

其中，$\tilde{Y}$ 为真实标签序列，Y_X 为所有可能的标注序列，$Score(X,Y)$ 为预测正确路径的得分，$Score(X,\tilde{Y})$ 为所有路径中某一条路径的得分。

解码时，使用维特比(Viterbi)算法获得得分最高的标签序列 y^*，其计算式如式(10)所示：

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

通过式(11)和式(12)可分别计算得到损失函数 L_{NER} 和 L_{CWS}，对 NER 任务和 CWS 任务中的训练样本分别进行训练并优化，从而最小化损失函数。

$$L_{\mathrm{NER}}=-\log P(Y\mid X) \tag{11}$$

$$L_{\mathrm{CWS}}=-\log P(Y'\mid X') \tag{12}$$

3.2 对抗训练

受 GAN 启发的对抗技术，将对抗训练纳入共享空间，以确保共享空间中不存在任务的私有信息。将由共享 BiLSTM 构成的共享特征提取器获得的 NER 任务和 CWS 任务的共享词边界信息输入最大池化层，去除冗余信息，进行特征压缩。将池化后的特征向量进行 sotfmax 分类，判断特征归属。任务鉴别器的表达式如式(13)和式(14)所示：

$$s=Maxpooling(h^s) \tag{13}$$

$$D(s;\delta_d)=\mathrm{softmax}(\boldsymbol{W}_d s+\boldsymbol{b}_d) \tag{14}$$

其中，h^s 为共享 BiLSTM 的特征输出，δ_d 为任务鉴别器的参数，$\boldsymbol{W}_d$ 为权重，$\boldsymbol{b}_d$ 为偏置项。

为了防止 CWS 任务的私有信息进入共享空间，引入对抗损失函数 L_{Adv} 训练共享特征提取器产生的共享特征，使任务鉴别器无法对来自 NER 任务和 CWS 任务的特征进行判别。对抗损失函数的计算式如式(15)所示：

$$L_{\mathrm{Adv}}=\min_{\delta_s}\left(\max_{\delta_d}\sum_{k=1}^{K}\sum_{i=1}^{T_k}\log D(E_s(x_k^{(i)}))\right) \tag{15}$$

其中，δ_s 为共享特征提取器中的可训练参数，E_s 为共享特征提取器，K 为任务的数量，T_k 为任务 k 中的训练样例数，$x_k^{(i)}$

为任务 k 中的第 i 个样例。

对于极大极小优化问题，给定一个句子，共享特征提取器生成一个表示来误导任务鉴别器，任务鉴别器尽可能对任务类型进行正确的分类。在训练过程中，共享特征提取器和任务鉴别器达到一个平衡点，即两者均无法继续改进且任务鉴别器也无法区分特征是来自 NER 任务还是 CWS 任务。最终将提取两个任务的共享词边界信息与 NER 任务的私有信息共同训练，以提高 NER 任务的准确率。对抗训练任务的伪代码如算法 1 所示。

算法 1 对抗训练任务

输入：NER 任务数据集和 CWS 任务数据集

输出：NER 任务和 CWS 任务的共享词边界信息

1. 输入数据，使用 MacBERT 获得 NER 任务和 CWS 任务的输入文本的向量表示
2. 将 NER 任务和 CWS 任务的向量表示输入共享 BiLSTM
3. For 使用对抗训练来提取 NER 任务和 CWS 任务的共享词边界信息
4. 共享 BiLSTM 进行特征提取
5. 由 Maxpooling 和 softmax 构成的任务鉴别器进行判别
6. 计算对抗训练的损失函数 $L_{Adv}=\min\limits_{\delta_s}(\max\limits_{\delta_d}\sum\limits_{k=1}^{K}\sum\limits_{i=1}^{T_k}\log D(E_s(x_k^{(i)})))$
7. 优化参数$[\delta_d, \mathbf{W}_d, \mathbf{b}_d, \delta_s]$，直到结果最优
8. End For
9. 输出 NER 任务和 CWS 任务的共享词边界信息

3.3 模型训练

通过对 NER 任务损失函数 L_{NER}、CWS 任务损失函数 L_{CWS} 和对抗损失函数 L_{Adv} 的计算，总损失函数 L 的计算式如式(16)所示：

$$L=GL_{NER}+(1-G)L_{CWS}+\gamma L_{Adv} \tag{16}$$

其中，γ 为损失权重系数，G 为判定输入是来自 NER 任务还是 CWS 任务的切换函数。

在模型训练过程中，从给定任务中每次抽取一个训练样例进行参数更新，不断优化最终的损失函数，并以 NER 任务的收敛速度为准进行迭代，直到结果最优。

4 实验结果与分析

4.1 实验数据

本文使用审计文本(Audit 2022)数据集用于 NER 任务的实验数据，在 SIGHAN 2006 数据集[18]上进行 NER 任务以验证模型在其他领域的泛化能力，将人民日报数据集[1)]用于 CWS 任务。数据集信息如表 2 所列。

表 2 各数据集句子数和实体数信息

Table 2 Information of the number of sentences and entities in each dataset

数据集	类型	训练集	验证集	测试集
Audit 2022	句子数	4603	1586	1500
	实体数	14647	4902	4533
SIGHAN 2006	句子数	46364	—	4365
	实体数	74703	—	6181
人民日报	句子数	20864	2318	4636
	实体数	33992	3819	7707

本文对审计文本数据集进行构建，并将其命名为 Audit 2022。过程如下：

(1)使用网络爬虫技术从中华人民共和国财政部官网爬取官方新闻报道审计文本数据共计 5395 篇。

(2)对爬取的文本按大小进行排序，保留 5 kB 以下的文本；对保留的文本进行预处理，包括去除非正文部分，编码统一等；以标点符号“。”“！”“？”进行分句；将所有的句子按照 6∶2∶2 的比例划分训练集、验证集和测试集，并对 4 类实体类型(人名、地名、机构名和专有名词)采用 BIO 方式(B 表示实体的开头、I 表示实体的中间或结尾、O 表示非实体)进行标注，共得到 9 种类型标注实体，分别为：B-PER，I-PER，B-LOC，I-LOC，B-ORG，I-ORG，B-AUD，I-AUD，O。

(3)使用我们构建的实体标注工具软件[2)]，对训练集和验证集中的句子进行实体标注和人工核对，可以大大节省人工标注成本。

(4)对测试集进行人工标注，在标注过程中，注意复合实体边界的划分及标注，标注完成后由专业人员进行核对。

人民日报数据集和 SIGHAN 2006 数据集分别包含 3 类实体类型，即人名、地名和机构名。

4.2 实验设置

本文实验环境如下：使用 Ubuntu 20.04 操作系统，显卡为 RTX3090，显存大小为 64 GB，Python 版本为 3.7.15，深度学习框架使用 PyTorch 版本为 1.9.0+cu111。

使用十折交叉验证对模型进行训练，具体超参设置如表 3 所列。

表 3 超参设置

Table 3 Hyperparameter settings

名称	值
embedding_dim	768
LSTM_dim	200
优化器	Adam
γ	0.06
batch_size	32
epoch	50
learning_rate	5×10^{-4}
dropout	0.5

4.3 实验指标

考虑本文各数据分布的不平衡性，采用精确率(Precision，P)、召回率(Recall，R)和微平均 F1 值(F1-micro)用于评价模型性能的指标，计算式如式(17)－式(19)所示：

$$P=\frac{TP}{TP+FP}\times100\% \tag{17}$$

$$R=\frac{TP}{TP+FN}\times100\% \tag{18}$$

$$F1\text{-}micro=\frac{2\times P_{micro}\times R_{micro}}{P_{micro}+R_{micro}}\times100\% \tag{19}$$

其中，TP 为正确样本被判断为正的数量，FP 为错误样本被判断为正的数量，FN 为正确样本被判断为负的数量，P_{micro} 为 P 所有类别的和，R_{micro} 为 R 所有类别的和。

4.4 实验结果与分析

4.4.1 嵌入方式的对比实验与分析

为了验证 MacBERT 中文预训练语言模型在 Audit 2022 数据集上的有效性，与 Word2vec，BERT 嵌入方式进行对比

1) https://github.com/OYE93/Chinese-NLP-Corpus

2) 软件著作权：《NER 实体标注工具软件[简称：NER-Tool] V1.0》

实验。实验结果如表 4 所列。

表 4　嵌入方式的对比

Table 4　Comparison of embedding methods

（单位：%）

模型	*P*	*R*	*F*1-*micro*
Word2vec＋BiLSTM＋CRF	74.78	86.93	80.40
BERT＋BiLSTM＋CRF	82.81	90.57	86.52
MacBERT＋BiLSTM＋CRF	85.42	91.20	88.22

由表 4 中的实验结果可知，使用 MacBERT 嵌入方式的模型 *F*1-*micro* 值为 88.22%，效果优于 Word2vec 和 BERT。相比使用 Word2vec 的模型，*P*，*R* 和 *F*1-*micro* 值分别提升了 8.03%，4.27%和 7.82%，这是因为 MacBERT 可以获得动态的向量表示，增强了对语义的理解；相比使用 BERT 的模型，*P*，*R* 和 *F*1-*micro* 值分别提升了 2.61%，0.63%和1.7%，这是因为 MacBERT 以中文为基础进行预训练，并且改进了 BERT 的两个预训练任务，其中 MLM 预训练任务可以有效缩小“预训练-下游任务”的差异。验证了 MacBERT 中文预训练语言模型的有效性。

4.4.2　对抗训练的对比实验与分析

为了验证对抗训练在 Audit 2022 数据集上的有效性，在 MacBERT 中文预训练语言模型的实验基础上加入对抗训练进行对比实验。实验结果如表 5 所列。

表 5　对抗训练的对比

Table 5　Comparison of adversarial training

（单位：%）

模型	*P*	*R*	*F*1-*micro*
MacBERT＋BiLSTM＋CRF	85.42	91.20	88.22
Audit-MBCA	91.10	91.00	91.05

由表 5 中的实验结果可知，相比未使用对抗训练的 MacBERT＋BiLSTM＋CRF 模型，Audit-MBCA 模型的 *F*1-*micro* 值提升了 2.83%，这是因为对抗训练的引入有效利用了 NER 任务和 CWS 任务的共享词边界信息帮助进行实体边界识别，因此 Audit-MBCA 模型效果最好。

4.4.3　泛化性能验证的对比实验与分析

为了验证模型在其他领域的泛化能力，在 SIGHAN 2006 数据集上与其他模型进行对比实验。实验结果如表 6 所列。

表 6　在 SIGHAN 2006 数据集上的对比

Table 6　Comparison on SIGHAN 2006 dataset

（单位：%）

模型	*P*	*R*	*F*1
Ensemble-SVM[19]	91.67	89.26	90.45
CNN-BiLSTM-CRF[20]	91.63	90.56	91.09
门控 CNN-CRF[21]	91.05	89.93	90.49
BiLSTM＋CRF＋adversarial＋self-attention[16]	91.73	89.58	90.64
DeepCAN(8)＋CRF＋adversarial[17]	93.55	90.15	91.82
Lattice＋LSTM-CRF[22]	93.57	92.79	93.18
BERT＋DeepCAN＋CRF[23]	93.82	92.24	93.37
Audit-MBCA	94.39	93.01	93.70

从表 6 中的结果可以得到以下结论：1）Ensemble-SVM 模型和 CNN-BiLSTM-CRF 模型主要通过融合词或字形来增强字符表示。与它们相比，本文模型的 *F*1 值分别提升了 3.25%和 2.61%，这是因为 MacBERT 通过预训练可以获得动态的向量表示，因此能够解决表征单一的问题。2）门控 CNN-CRF 模型通过单词嵌入加位置嵌入获得向量表示，并使用门控 CNN 代替 BiLSTM 进行特征提取。与本文模型相比，其未能使用预训练语言模型和考虑词边界信息。3）BiLSTM＋CRF＋adversarial＋self-attention 模型和 DeepCAN(8)＋CRF＋adversarial 模型均从预先训练好的字符嵌入矩阵中获取嵌入向量，并引入对抗训练提取两个任务共享的词边界信息。与这两个模型相比，本文模型 *F*1 值分别提升了 3.06%和 1.88%，说明 MacBERT 获得中文数据集的表示效果更好。4）Lattice＋LSTM-CRF 模型将所有潜在的分词信息和字符信息进行编码，取得了 93.18%的 *F*1 值，说明分词对 NER 任务尤为重要。与本文模型相比，其未能使用对抗训练方法，因而所有潜在的分词信息可能会给 NER 任务带来噪声。5）BERT＋DeepCAN＋CRF 模型通过叠加多层卷积注意力模块构建强特征器，该强特征器能够提取局部特征信息和句子层级特征。与之相比，本文模型 *F*1 值提升了 0.33%，这是因为 BERT 存在“预训练-下游任务”不一致的问题，且在特征提取时未能考虑词边界信息。

与上述模型对比，本文提出的 Audit-MBCA 模型性能更好，*F*1 值提升了 0.33%～3.25%，验证了本文模型的泛化性能。

结束语　本文针对审计文本命名实体识别任务中缺乏审计文本语料库和实体边界识别不清晰的问题，提出了基于 MacBERT 和对抗训练的审计文本命名实体识别模型（Audit-MBCA）。首先构建了审计文本数据集并将其命名为 Audit 2022；其次通过 MacBERT 中文预训练语言模型获得审计文本更好的向量表示；最后引入对抗训练，提取 NER 任务和 CWS 任务的共享词边界信息，并与 NER 任务的私有信息共同训练，以提高 NER 任务的准确率。通过实验验证了本文模型的有效性。在未来的研究工作中，将从两方面入手：一是对审计文本数据集进行进一步的扩充；二是通过实体识别与实体关系构建政策跟踪审计系统。

参 考 文 献

[1] ZHANG W，WU Z A. Application of Natural Language Analysis of Unstructured Text Data in Policy Tracking Audit[J]. Audit Observation，2022(4)：70-75.

[2] CHEN X，OUYANG C，LIU Y，et al. Improving the named entity recognition of Chinese electronic medical records by combining domain dictionary and rules[J]. International Journal of Environmental Research and Public Health，2020，17(8)：2687-2703.

[3] YU H K，ZHANG H P，LIU Q，et al. Chinese named entity identification using cascaded hidden Markov model[J]. Journal on Communications，2006，27(2)：87-94.

[4] ZHANG Y J，XU Z T，XUE X Y. Fusion of Multiple Features for Chinese Named Entity Recognition Based on Maximum Entropy Model[J]. Journal of Computer Research and Development，2008，45(6)：1004-1010.

[5] TANG B Z，CAO H X，WU Y H，et al. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features[J]. BMC Medical Informatics and Decision Making，2013，13(S1)：1-10.

[6] PATIL N，PATIL A，PAWAR B V. Named entity recognition using conditional random fields[J]. Procedia Computer Science，

2020, 167: 1181-1188.
[7] HAMMERTON J. Named entity recognition with long short-term memory[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. 2003: 172-175.
[8] LAMPLE G, BALLESTEROS M, SUBRA-MANIAN S, et al. Neural architectures for named entity recognition[J]. arXiv: 1603.01360, 2016.
[9] CHI Y N. Research on Question and Answer Technology of Corporate Financial Audit Based on Deep Learning[D]. Harbin: Harbin Engineering University, 2018.
[10] CUI Y, CHE W, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[J]. arXiv: 2004.13922, 2020.
[11] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.
[12] ZHANG H F, ZENG C, PAN L. News topic text classification method based on BERT and feature projection network[J]. Journal of Computer Applications, 2022, 42(4): 1116-1124.
[13] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese bert[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.
[14] JIAO K N, LI X, YE H, et al. Fine-grained entity recognition based on MacBERT-BiLSTM-CRF in anti-terrorism field[J]. Science Technology and Engineering, 2021, 21(29): 12638-12648.
[15] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Nets[C]// Neural Information Processing Systems. MIT Press, 2014: 2672-2680.
[16] CAO P, CHEN Y, LIU K, et al. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 182-192.
[17] ZHANG L L. Research on Identification of the Chinese Named Entity Based on Deep Learning[D]. Taiyuan: Taiyuan University of Science and Technology, 2021.
[18] LEVOW G A. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition[C]// Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. 2006: 108-117.
[19] YIN Z Z, LI X Z, HUANG D G, et al. Chinese Named Entity Recognition Ensembled with Character[J]. Journal of Chinese Information Processing, 2019, 33(11): 95-100, 106.
[20] JIA Y, XU X. Chinese named entity recognition based on CNN-BiLSTM-CRF[C]// 2018 IEEE 9th International Conference on Software Engineering and Service Science(ICSESS). IEEE, 2018: 1-4.
[21] TAO Y, PENG Y B. Chinese named entity recognition based on Gated-CNN-CRF[J]. Electronic Design Engineering, 2020, 28(4): 42-46, 51.
[22] ZHANG Y, YANG J. Chinese NER using lattice LSTM[J]. arXiv: 1805.02023, 2018.
[23] XIE B H, ZHANG L L, ZHAO H Y. Chinese Named Entity Revognition Method Based on BERT-DeepCAN-CRF[J]. Computer & Digital Engineering, 2022, 50(12): 2720-2726.

QIAN Taiyu, born in 1994, postgraduate, is a member of China Computer Federation. His main research interest is text mining.

CHEN Yifei, born in 1977, Ph.D, associate professor. Her main research interests include text mining and intelligent information extraction.