Journal of Computer Applications
计算机应用
ISSN 1001-9081
CODEN JYIIDU
http://www.joca

doi:10.11772/j.issn.1001-9081.2021071209

基于知识库实体增强 BERT 模型的中文命名实体识别

胡婕[1*]，胡燕[1]，刘梦赤[2]，张龑[1]

(1.湖北大学 计算机与信息工程学院，武汉 430062；

2. 华南师范大学 计算机学院，广州 510898)

(*通信作者电子邮箱 JieHu@hubu.edu.cn)

摘 要: 针对预训练模型 BERT 存在词汇信息缺乏的问题，在半监督实体增强最小均方差预训练模型的基础上提出了一种基于知识库实体增强 BERT 模型的中文命名实体识别模型 OpenKG + Entity Enhanced BERT+CRF。首先从中文通用百科知识库 CN-DBPedia 中下载文档并用结巴分词抽取实体来扩充实体词典，然后将词典中的实体嵌入到 BERT 中进行预训练，将训练得到的词向量输入到双向长短期记忆(BiLSTM)网络中提取特征，最后经过条件随机场(CRF)修正后输出。在 CLUENER 2020 和 MSRA 数据集上进行模型验证，将所提模型分别与 Entity Enhanced BERT Pre-training、BERT+BiLSTM、ERNIE 和 BiLSTM+CRF 模型进行对比实验，该模型的 F1 值在两个数据集上相比于对比模型分别提高了 1.63 和 1.1 个百分点，3.93 和 5.35 个百分点，2.42 和 4.63 个百分点以及 6.79 和 7.55 个百分点。实验结果表明，所提模型对命名实体识别的综合效果得到有效提升，F1 值均优于对比模型。

关键词: 命名实体识别；知识库；实体词典；预训练模型；双向长短期记忆网络

中图分类号: TP 391.1　　**文献标志码:** A

Chinese named entity recognition based on knowledge base and entity enhanced BERT model

HU Jie[1*], HU Yan[1], LIU Mengchi[2], ZHANG Yan[1]

(1. School of Computer Science and Information Engineering, Hubei University, Wuhan Hubei 430062, China;

2. School of Computer Science, South China Normal University, Guangzhou Guangdong 510898, China)

Abstract: Aiming at the problem that the pre-training model BERT lacks of vocabulary information, a Chinese named entity recognition model called Open knowledge base +Entity Enhanced Bidirectional Encoder Representation from Transformers + Conditional Random Fields (OpenKG + Entity Enhanced BERT + CRF) based on open domain knowledge base entity enhanced Bert model was proposed in terms of the idea of the semi-supervised entity-enhanced minimum mean-square error pre-training model. Firstly, documents were downloaded from CN-DBPedia and entities were extracted by Jieba Chinese text segmentation to expand entity dictionary, and then the entities in the dictionary were embedded into BERT for pre-training. The word vectors obtained from the training were input into Bidirectional Long-Short-Term Memory (BiLSTM) network for feature extraction. Finally, the output was corrected by Conditional Random Field (CRF). Model validation was performed on data sets CLUENER 2020 and MSRA, and the proposed model was compared with Entity Enhanced BERT pre-training, BERT+BiLSTM, ERNIE and BiLSTM+CRF models. Compared with the comparison models, the F1 score of the proposed model increased by 1.63 and 1.1 percentage points, 3.93 and 5.35 percentage points, 2.42 and 4.63 percentage points, 6.79 and 7.55 percentage points respectively in the two data sets. Experimental results show that the comprehensive effect of the proposed model on named entity recognition is effectively improved, and the F1 scores are better than those of the comparison models.

Keywords: named entity recognition; knowledge base; entity dictionary; pre-training model; Bidirectional Long Short-Term Memory(BiLSTM) network

收稿日期: 2021-07-12;　**修回日期**:2021-09-18;　**录用日期**:2021-09-24。

基金项目: 国家自然科学基金资助项目（61977021）；广州市大数据与智能教育重点实验室资助项目（201905010009）

作者简介: 胡婕(1977—)，女，湖北汉川人，教授，博士，主要研究方向：复杂语义大数据管理、自然语言处理；胡燕(1993—)，女，安徽安庆人，硕士研究生，主要研究方向：自然语言处理；刘梦赤(1962—)，男，湖北武汉人，教授，博士，CCF 会员，主要研究方向：语义数据库、深度学习；张龑（1974—），男，湖北宜昌人，教授，博士，CCF 会员，主要研究方向：软件工程、信息安全。

0 引言

命名实体识别（Named Entity Recognition，NER）在自然语言处理（Natural Language Processing, NLP）的许多下游任务如知识库构建[1]，信息检索[2]，以及问答系统[3]中扮演着重要角色。NER 任务主要方法有 3 种：基于规则、基于传统机器学习和基于深度学习的方法。其中基于深度学习的方法与基于规则和基于统计学的方法相比无需人工设置特征，神经网络可以自动从数据中学习特征，被广泛的应用于命名实体识别的任务中[4-6]。

NER 任务传统上被解决为序列标记问题[7]，中文命名实体识别任务与英文有所不同，英文句子中有天然的边界，而中文句子没有，这给中文命名实体识别带来了更多挑战。因此中文 NER 任务在进行单词序列标注之前，一般要先进行分词然后再将词级别的序列标注模型应用于所分割的句子，命名实体边界也就是词的边界。然而，分词不可避免地会出现单词的错误划分从而造成实体边界的错误识别。为了解决分词错误对命名实体识别任务的影响，Li 等[8]认为基于字符的方法比基于词的方法更适用于中文 NER 任务，因为它没有中文分词错误。然而，基于字符的 NER 任务并不能充分利用词和词的序列信息。为了解决这个问题，Zhang 等[9]提出了格结构的长短期记忆网络（Lattice Long Short-Term Memory network, Lattice-LSTM）模型，将词典纳入基于字符的模型。此外，当字符与词典中的多个单词匹配时，保留所有与字符匹配的单词，而不是启发式地为字符选择一个单词，让后续的 NER 模型来确定应用哪个单词。通过使用 Lattice-LSTM 来表示句子中的词汇，将潜在的单词信息整合到基于字符的 LSTM-CRF(Long Short-Term Memory network-Conditional Random Fields)中，但是 Lattice-LSTM 的模型架构相当复杂。为了引入词典信息，Lattice-LSTM 在输入序列中不相邻的字符之间添加了多个额外的边，显著降低了其训练和推理速度，而且很难将网格模型的结构应用到其他神经网络结构，只适合于某些特定的任务，不具通用性。于是 Ma 等[10]提出更简单的方法来实现 Lattice-LSTM 的思想，将每个字符的所有匹配单词合并到基于字符的 NER 模型中，在字符中表示编码词典信息，并设计编码方案以尽可能多地保留词典匹配结果。这种方法不需要复杂的模型结构，更容易实现，并且可以通过调整字符表示层快速适应任何合适的神经 NER 模型。然而由于网格结构的复杂性和动态性，现有基于网格的模型很难充分利用图形处理器进行并行计算，因而推理的速度通常较慢。因此，Li 等[11]提出了平面点阵变换器，核心是将点阵结构转换成一组跨度，并引入特定的位置编码，在性能和效率上优于其他基于词典的模型。Xue 等[12]和 Gui 等[13-14]利用词汇特征，外部词汇级信息增强了 NER 训练。

然而，上述方法都是有监督的模型，当处理有较少标记数据的数据集时，小数据无法反映出语言间的复杂关系，同样也很容易让复杂的深度网络模型过拟合，很难获得很好的训练网络。因此预先训练的半监督语言模型就显得尤为重要。Devlin 等[15]提出的 BERT(Bidirectional Encoder Representations from Transformers)模型就是一个预训练半监督模型，可以在与最终任务无关的大数据集上训练出语言的表示，然后将学到的知识表示用到任务相关的语言表示上。Sun 等[16]提出了 ERNIE(Enhanced Language Representation with Informative Entities)模型，它通过知识整合来增强 BERT。ERNIE 通过屏蔽完整实体来训练，而不是像 BERT 那样屏蔽单个字词标记。ERNIE 预训练的实体级掩码技巧可以看作是一种通过错误反向传播来集成实体信息的隐式方法。由于命名实体识别中的实体可能出现二义性，即相同的词在不同的领域有不同的语义，因此包含领域的实体词典对于该任务是有用的。考虑到这一点，Jia 等[17]提出了将词典嵌入到针对中文 NER 的预先训练最小均方误差模型中，提出了一种半监督实体增强的最小均方误差预训练模型 Entity Enhanced BERT Pre-training。具体来说，首先使用新词发现方法从原始文本以及相关文档中提取实体词典。然后使用 Char-Entity-Self-Attention 机制替换原始的自我注意力机制将实体信息嵌入到 BERT 中，也就是使用字符和实体表示组合来增强自我关注。该机制可以更好地捕捉字符和文档特定实体的上下文相似性，并将字符隐藏状态与每一层中的实体嵌入显示结合。但是此方法提取实体词典的方式较为复杂而且获取的实体词数量和使用范围有限。如今，开放域和领域知识库构建越来越完善，可免费获得的知识库也越来越多。因此本文提出了在词典中加入知识库信息的方法来扩展词典中的实体信息，使词典中的词使用更具广泛性。具体来说，首先在中文通用百科知识图谱 CN-DBpedia[18]中下载其提供的 mention2entity 文档，该文档中包含了 110 多万条数据，这些数据中包含了大量的实体，使用结巴分词对数据进行分词处理，留下带有名词标签的词，使得词典中的实体词更丰富，应用领域更广泛。而且由于各个领域的实体词典可以从其领域知识库中获得，可以减少前期词典创建的工作量。随后将词典中的实体嵌入到 BERT 预训模型中进行预训练，然后在 NER 微调任务中将训练得到的词向量输入到 BiLSTM 中提取特征，最后通过 CRF 层从训练数据中获得约束性规则，为最后预测的标签添加约束来保证预测标签的合法性，生成最优序列结果。实验结果表明本文提出的模型在 CLUENER 2020[19]数据集上的 F1 值达到了 78.15%，在 MSRA[20]数据集上的 F1 的值达到了 88.11%，相比于上述 Entity Enhanced BERT Pre-training 模型以及其他三个基线模型 BERT+BiLSTM、ERNIE 和 BiLSTM + CRF 都有所提升，从而验证了加入知识库之后的词典结构在中文 NER 语言模型预训练中整合实体信息的有效性，以及在实体识别的微调任务中加入 CRF 层预测标签的有效性。

1 本文方法

本文的命名实体识别方法主要分为三个部分首先从中文通用百科知识库 CN-DBPedia 中抽取实体来构建实体词典，然后将词典中的实体嵌入到 BERT 中进行预训练，将训练得到的词向量输入到 BiLSTM 提取特征，最后经过条件随机场修正后输出。

1.1 词典的构建

为了获得特定文档的实体信息，将其嵌入到 BERT 预训练语言模型中，Jia 等采用 Bouma 等[21]所提出的无监督方法在原始文档中自动发现候选实体，分别计算连续字符之间的交互信息值和左、右熵度量值，然后将这三个值相加作为可能实体的有效评分。

本文在此基础上加入开放域知识库中所提供的实体来对原有的词典进行扩充，将实体词典扩充成一个大小为 6086KB 的实体词典。本文使用的知识库是由复旦大学知识工场实验室研发并维护的大规模通用百科知识图谱知识库 CN-DBpedia，其数据来源于中文百科类网站如百度百科、互动百科、中文维基百科等的纯文本页面中提取的信息，经过过滤、融合、推断等处理后，最终形成的高质量结构化数据。本文使用 CN-DBPedia 所提供 mention2entity 文档中的数据，其包含 110 多万条信息，包含了大量的实体，所包含的领域非常广泛，获取的途径也很方便。本文的具体做法，从 OpenKG.CN 网站下载 mention2entity 中的文本后对数据进行清洗，清洗的过程是用可以标注词性的“结巴”分词工具对文本进行全模式分词，将标注为名词词性的词挑选出来，去掉重复的词语，将剩余的词加入到词典中作为候选实体。

1.2 嵌入词典实体信息的 BERT 预训练

嵌入实体信息的 BERT 预训练模型结构如图 1 所示，与基于中文 BERT[15]的 Transformer 模型中 Encoder 结构类似，为了利用提取的实体，即将实体信息嵌入到模型中，将 Transformer 扩展为 Char-Entity-Transformer，如图 2 所示它是由一个多头的 Char-Entity-Self-Attention 块堆栈组成。

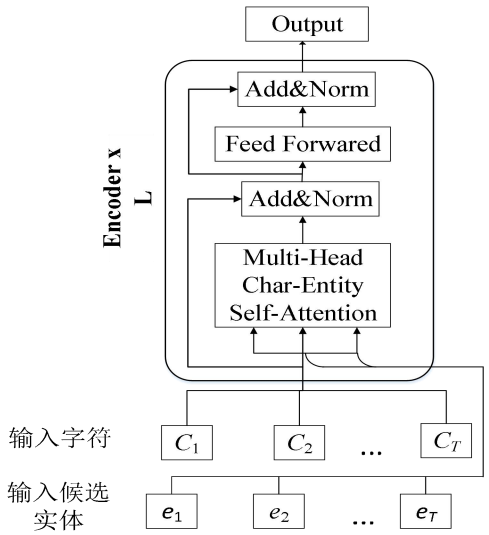

图 1 嵌入实体信息的 BERT 预训练模型结构

Fig. 1 Structure of BERT model embedding entity information

首先将字符与提取的实体进行匹配，给定字符序列 $c=\{c_1,c_2...,c_T\}$ 和提取的实体字典 Entity，使用最大实体匹配算法得到对应的实体标记序列 $e=\{e_1,e_2...,e_T\}$。用包含该字符词典中最长实体的索引来标记每个字符，并将没有实体匹配的字符标记为 O。

在模型的输入阶段，给定一个字符序列 $c=\{c_1,c_2...,c_T\}$，输入层中的第 t 个字符的表示是字符，文本和位置嵌入的总和，表示为：

$$\boldsymbol{h}_t^l=\boldsymbol{E}_c[c_t]+\boldsymbol{E}_s[0]+\boldsymbol{E}_p[t] \tag{1}$$

其中 $\boldsymbol{E}_c$，$\boldsymbol{E}_s$，$\boldsymbol{E}_p$ 分别表示字符的字嵌入查找表，文本嵌入查找表和位置查找表，因为没有用到下一句预测任务的输入句子顺序，所以将文本索引 s 设置为常数 0。

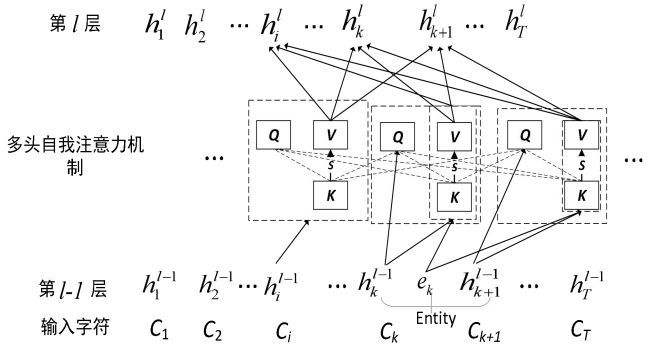

图 2 Char-Entity-Self-Attention 模型结构图

Fig. 2 Structure of Char-Entity-Self-Attention model

接下来将给定字符序列和前面所得到的实体标记序列一起输入到如图 2 所示的多头 Char-Entity-Self-Attention 模型，将汉字的隐含维数和新词实体的隐含维数分别表示为 $\boldsymbol{H}_C$ 和

$\boldsymbol{H}_e$，L 是层数，A 是自我关注头数。对于给定（l-1）层字符的隐藏序列 $\{h_1^{l-1}, h_2^{l-1}..., h_T^{l-1}\}$ 这里的 Key 矩阵和 Value 矩阵与 BERT 的多头注意力有所不同，这里用实体的隐藏字符和实体嵌入组合来生成 key 和 Value 矩阵，表示为：

$$q_t^l = h_t^{l-1\top} \boldsymbol{W}_{h,q}^l \tag{2}$$

$$\mathrm{k}_t^l = \begin{cases} h_t^{l-1\top} \boldsymbol{W}_{h,k}^l & \text{if } \mathrm{e}_t = 0, \\ \frac{1}{2}\left(h_t^{l-1\top} \boldsymbol{W}_{h,k}^l + \boldsymbol{E}_{ent}^{\top}[e_t] \boldsymbol{W}_{e,k}^l\right) & \text{else;} \end{cases} \tag{3}$$

$$v_t^l = \begin{cases} h_t^{l-1\top} \boldsymbol{W}_{h,v}^l & \text{if } \mathrm{e}_t = 0, \\ \frac{1}{2}\left(h_t^{l-1\top} \boldsymbol{W}_{h,v}^l + \boldsymbol{E}_{ent}^{\top}[e_t] \boldsymbol{W}_{e,v}^l\right) & \text{else;} \end{cases} \tag{4}$$

这里的 $\boldsymbol{Q}$ 矩阵与 BERT 的相同。其中 $\boldsymbol{W}_{h,q}^l, \boldsymbol{W}_{h,k}^l, \boldsymbol{W}_{h,v}^l \in \boldsymbol{R}^{H_c * H_c}$ 是第 l 层的可训练参数，$\boldsymbol{W}_{e,k}^l, \boldsymbol{W}_{e,v}^l \in \boldsymbol{R}^{H_e * H_c}$ 是对应实体可训练参数，$\boldsymbol{E}_{ent}$ 是实体嵌入查找表。将一组实体嵌入 $\{\boldsymbol{E}_{ent}[e_1], \boldsymbol{E}_{ent}[e_2]...\boldsymbol{E}_{ent}[e_T]\}$ 表示为 $e \in \boldsymbol{R}^{T * \boldsymbol{H}_e}$，第 l 层第 i 个字符的注意力得分 S_i^l 计算为：

$$\begin{aligned} S_i^l &= \mathrm{softmax}\left\{\frac{q_i^l \boldsymbol{K}^{l\top}}{\sqrt{d_k}}\right\} \\ &= \mathrm{softmax}\left\{\frac{q_i^l \left(h^{l-1} \boldsymbol{W}_{h,k}^l + e \boldsymbol{W}_{e,k}^l\right)^{\top}}{2\sqrt{d_k}}\right\} \\ &= \left\{\frac{\sqrt{s_t^c s_t^e}}{\sum_j \sqrt{s_j^c s_j^e}}\right\}_{t=1}^{\mathrm{T}} \end{aligned} \tag{5}$$

$$s.t.\ \ \mathrm{s}_t^c = \exp\left(\frac{q_i^l \left(h_t^{l-1\top} \boldsymbol{W}_{h,k}^l\right)^{\top}}{\sqrt{d_k}}\right); \tag{6}$$

$$\mathrm{s}_t^e = \exp\left(\frac{q_i^l \left(e_t^{\top} \boldsymbol{W}_{e,k}^l\right)^{\top}}{\sqrt{d_k}}\right)$$

其中，字符到字符之间的注意力分数 s_t^c 的计算与基线 BERT 的自我注意力相同。字符到实体注意力分数 s_t^e 表示字符与相应实体之间的相似性。归一化前，第 i 个字符和第 t 个字符的注意力得分 $\{S_i^l\}_t$ 是 $\sqrt{\mathrm{s}_t^c \mathrm{s}_t^e}$，即 s_t^c 和 s_t^e 的几何平均值。这表明通过 Char-Entity-Self-Attention 计算两个字符之间的相似度是作为字符到字符几何距离和字符到实体几何距离的组合计算。给定注意力分数 S_i^l，$\boldsymbol{V}^l$ 为字符值和实体值的组合，则 $Atten\left(q_i^l, \boldsymbol{K}^l, \boldsymbol{V}^l\right)$ 计算如公式(7)所示：

$$\begin{aligned} Atten\left(q_i^l, \boldsymbol{K}^l, \boldsymbol{V}^l\right) &= \boldsymbol{S}_i^l \boldsymbol{V}^l \\ &= \boldsymbol{S}_i^l \frac{1}{2}\left(h^{l-1} \boldsymbol{W}_{h,v}^l + e \boldsymbol{W}_{e,v}^l\right) \end{aligned} \tag{7}$$

1.3 NER 任务

本文 NER 任务模型框架如图 3 所示。

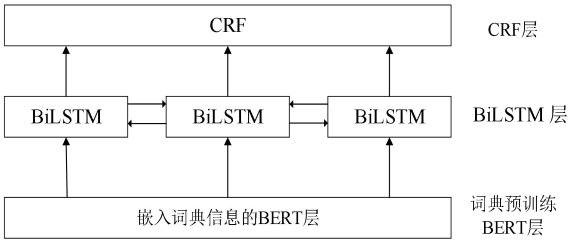

图 3 NER 任务模型框架图

Fig. 3 NER task model framework diagram

将文本信息转化为计算机可以识别的数据形式是任务的第一步。目前常用的词嵌入模型主要是 BERT 预训练语言模型，它通过双向 Transformer 编码器生成字向量，但是实体识别的任务是识别人名，地名等实体信息，BERT 模型无法利用现有的实体信息。本文使用如 1.2 节所述的嵌入实体信息的 BERT 预训练模型，其 Char-Entity-Self-Attention 机制可以很好地捕捉字符和文档特定实体的上下文相似性，并显示地将字符隐藏状态与每一层实体嵌入结合，再将扩展后的模型对数据集信息进行编码。将嵌入词典实体预训练 BERT 模型的最后一层输出输入到 BiLSTM 中进行训练，进一步提取文本特征。通过 BiLSTM 对序列的上下文信息进行学习，为每个标签打分，BiLSTM 的输出为字符的每一个标签分值，输出结果如图 4 所示。

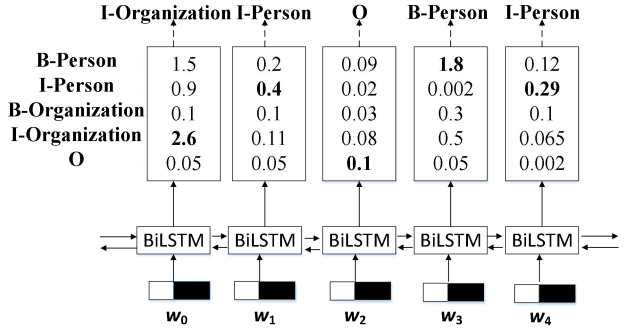

图 4 BiLSTM 输出标签

Fig. 4 BiLSTM output label

BiLSTM 通过挑选每个标签分值最高的作为该字符的标签，并选取最大分值作为每个字符最终的输出标签。如图 4 所示，模型所生成的标签为 I-Organization I-Person 和 B-Person I-Person，但是显然 I-Organization 之后不可能出现

I-Person，即：标签序列“I-Organization I-Person”是错误的。这种取最大值的方法虽然可以得到正确的标签序列B-Person I-Person，但是并不能保证每次的预测都是正确的，因此模型不能仅仅以 BiLSTM 的输出结果作为最终的预测标签，需要在预测标签与标签之间引入约束条件来保证生成标签的合法性，因此本文将 BiLSTM 的输出结果输入到 CRF 层。CRF 层可以为最后预测标签添加约束关系来保证预测标签的合理性。

给定输入序列 $\boldsymbol{X}=\{x_1,x_2,...,x_n\}$，假设训练得到对应输出标签序列 $\boldsymbol{Y}=\{y_1,y_2,...,y_n\}$。其中 n 代表 NER 标签的数量，则标签序列的得分可表示为：

$$s(\boldsymbol{X},y)=\sum_{i=1}\left(\boldsymbol{Z}_{y_i,y_{i+1}}+\boldsymbol{P}_{i+1,y_{i+1}}\right) \tag{8}$$

其中 $\boldsymbol{Z}$ 为转移矩阵；$\boldsymbol{Z}_{y_i,y_{i+1}}$ 为标签从 y_i 转移到 y_{i+1} 的分值；$\boldsymbol{P}_{i+1,y_{i+1}}$ 为输入序列第 $i+1$ 个字对应标签 y_{i+1} 的分值。对标签序列 y 的概率进行计算，可表示为：

$$P(y\mid\boldsymbol{X})=\frac{\exp\left(score\left(\boldsymbol{X},y\right)\right)}{\sum_{\tilde{y}\in Y_X}\exp\left(score\left(\boldsymbol{X},\tilde{y}\right)\right)} \tag{9}$$

其中 Y_X 为所有可能的标签序列集合，最终输出序列的标签为概率最大的标签集合。

2 实验结果与分析

2.1 数据集与评价指标

2.1.1数据集

为了验证本文模型的有效性，在两个公开使用的数据集 CLUENER 2020 和 MSRA 上做了对比实验。

数据集 CLUENER 2020[19]是一个细粒度的中文 NER 数据集，包含 10 种不同的实体类别，分别是 address, book, company, game, government, movie, organization, position, scene，并对常见类别进行了细粒度的划分，如将“组织”细分为“政府”和“公司”等。同时存在同一实体在不同语境下属于不同类别的情况，如 Twins 的字面意思是双胞胎，但是在娱乐新闻的背景下，它指的是 Twins 组合。本文从 CLUENER 2020 数据集中随机抽取 5200，600 和 748 条句子分别作为训练集，评估集和测试集，并将抽取的句子划分为四个新闻领域：GAM（游戏）、ENT（娱乐）、LOT（彩票）和 FIN（金融）。

MSRA[20]是中文 NER 的通用数据集。它包括三种类型的实体，分别是 PER（人名），LOC（地名），ORG（组织名）。本文使用标记集{B,I,E,O}进行标记。

数据集的详细信息如表 1 所示。

表 1 数据集描述

Tab. 1 Description of datasets

数据集	领域	CLUENER 2020		MSRA	
		句子	实体	句子	实体
Train	—	5200	10800	16000	32000
Dev	—	600	1200	—	—
Test	GAM(Game)	300	500	—	—
	ENT(Entertainment)	48	100	—	—
	LOT(Lottery)	100	300	—	—
	FIN(Finance)	300	600	—	—
	All	748	1500	4000	9000

2.1.2评估标准

本文采用准确率 P，召回率 R，和 F1 的值作为评价指标，这三种评价指标越高，代表模型的精确率，召回率和综合性能越好。评价指标的计算公式如下：

$$\text{准确率}=\frac{\text{正确识别的实体个数}}{\text{识别出实体总数}}\times100\%$$

$$\text{召回率}=\frac{\text{正确识别的实体个数}}{\text{标本标注的实体总数}}\times100\%$$

$$\text{F1}=2\times\frac{\text{准确率}\times\text{召回率}}{\text{准确率}+\text{召回率}}\times100\%$$

2.2 实验环境

本文所有的实验均在 Python3.6，pytorch1.7，GTX 5000 平台上运行。

2.3 实验设置

本文模型使用 BERT 构建，层数 $L=12$，自注意头数 $A=12$，字符的隐藏大小 $\boldsymbol{H}_C=768$，实体的隐藏大小 $\boldsymbol{H}_e=64$。其他超参数的设置如表 2 所示。

表 2 模型参数

Tab. 2 Parameters of our model

参数	BERT 预训练	NER 任务
Epoch	3	30
Batchsize	32	30
最大句子度	180	32
Optimize	Adam	-
学习率	3E-5	5E-5
Lr decay rate	0.01	-

2.4 实验结果与分析

（1）本模型与基线模型实验对比

本文对比的基线模型是 Entity Enhanced BERT Pre-training[17]。它首先在与数据集相关

的文档中获取词，将其作为候选实体放入实体词典中。然后将实体词典信息嵌入到BERT的预训练中，并将预训练模型用于NER任务中进行实体的分类输出。但是它在前期词典的获取过程中使用的方法并不能识别所提取词是否是真正的实体，导致词典中真正的命名实体比例降低，而加入了无关实体的词进行预训练会降低模型的性能。本文词典的提取方法是使用开放域知识库CN-DBpedia，其包含大量的实体三元组，用于提高抽取的候选实体中真正实体的比例。此外，基线模型在NER任务中没有利用CRF层来对生成的标签进行约束。

为了验证本文模型的有效性，在同一实验环境下，设计了两组实验与基线模型 Entity Enhanced BERT Pre-training[17]在测试集上的对比，在CLUENER 2020数据集和MSRA数据集上F1值的对比如表3所示。

表3 测试集上模型F1值(%)的对比

Tab. 3 Comparison of F1 scores (%) for each model on test sets

模型	CLUENER 2020					MSRA
	GAM	ENT	LOT	FIN	ALL	
文献[17]模型	70.90	87.11	82.73	77.18	76.52	87.01
OpenKG + 文献[17]模型	71.40	87.82	83.32	77.52	77.44	87.59
本文模型	**71.50**	**88.43**	**84.21**	**78.12**	**78.15**	**88.11**

从表3可知，与基线模型Entity Enhanced BERT Pre-training相比，本文加入开放域知识库的实体增强BERT模型OpenKG + Entity Enhanced BERT Pre-training在上述两个数据集上F1值都有一定的提升。从CLUENER 2020数据集的所有类别（All）F1的值可以看出，加入知识库之后的模型F1值提升了0.92个百分点，在MSRA数据集上F1值提升了0.58个百分点。这是因为本文选取的开放域知识库中mention2entity文档包含110多万条信息，包含了各个领域的大量实体，本文从中提取了三千多条候选词加入到对应新闻领域词典中。F1值的提升可以验证加入开放域知识库的有效性。在此基础上，本文使用OpenKG + Entity Enhanced BERT Pre-training+ CRF模型在NER微调中加入CRF层来修正标签，从表中CLUENER 2020数据集所有领域（All）的F1可以看出，相比于只加入开放域知识库的模型OpenKG + Entity Enhanced BERT Pre-training 的F1值提升了0.71个百分点，在MSRA数据集上F1值提升了0.52个百分点。相比于基线Entity Enhanced BERT Pre-training模型，在CLUENER 2020数据集上F1值提升了1.63个百分点，在MRSA数据集F1值提升了1.10个百分点，验证了NER微调加入CRF解码层的有效性。

Entity Enhanced BERT Pre-training模型与本文的两组模型在CLUENER 2020数据集的所有领域（All）和MSRA数据集的测试集上准确率，召回率和F1的值对比如表4所示。

表4 测试集上的模型对比

Tab. 4 Comparison of models on test sets

模型	CLUENER 2020			MSRA		
	准确率/%	召回率/%	F1值/%	准确率/%	召回率/%	F1值/%
文献[17]模型	74.45	78.53	76.52	86.04	88.02	87.01
OpenKG + 文献[17]模型	74.52	80.60	77.44	86.23	89.01	87.59
本文模型	**76.26**	**80.13**	**78.15**	**87.23**	**89.01**	**88.11**

从表4中可以看出，在CLUENER 2020和MSRA这两个公开的数据集上，本文模型在准确率、召回率和 F1 值上均有提升，验证了本文模型综合效果更佳。

（2）与相关工作对比

为了进一步验证本文模型的有效性，本文还对三组中文NER方法在CLUENER 2020数据集和MSRA数据集上进行了比较。这三组模型分别为：BERT+BiLSTM, ERNIE[22]以及BiLSTM+CRF模型。其中，ERNIE是百度公司基于BERT模型进一步优化得到的模型，它在中文NLP任务上获得了最佳效果。它主要是在掩码(mask)机制上做了改进，在预训练阶段不仅采取字掩码机制，而且增加了外部知识进一步采取全词掩码和实体掩码的三级掩码机制。

三组模型与本文模型在CLUENER 2020数据集的所有领域（All）和MSRA数据集F1值对比如表5所示。

表5 相关模型F1值(%)的对比

Tab. 5 Comparison of F1 scores (%) for related models

模型	CLUENER 2020	MSRA
BERT+BiLSTM	74.22	82.76
ERNIE	75.73	83.48
BiLSTM+CRF	71.36	80.56
本文模型	78.15	88.11

从表5可以看出，相比于直接对预先训练的中文BERT生成字向量与利用BiLSTM方法解码的模型，本文模型在CLUENER 2020数据集上的F1值提升了3.93个百分点，在MSRA数据集上的F1值提升了5.35个百分点，这是因为Char-Entity-Transformer结构能够有效地利用实体词典信息，并且考虑到不同实体在不同语境下可能有不同的语义的情况，利用CRF解码层为最后预测的标签添加约束关系来保证预测标签的合法性，从而提高了F1的值。与ERNIE相比，尽管ERNIE使用更多来自网络资源的原始文本和实体信息进行预训练，但是在CLUENER 2020数据集上F1值仍提升了2.42个百分点，并且在MSRA数据集上F1的值提升了4.63个百分点，这表明了通过字符-实体转换结构集成实体信息的显式方法比实体级掩蔽方法对中文NER更有效。与BiLSTM+CRF模型相比，在CLUENER 2020数据集上F1的值提升了6.79个百分点，在MSRA数据集上F1的值提升了7.55个百分点。这是因为嵌入词典实体的BERT预训练模型能够将实体集成到具

有字符实体转换器的结构中，从而提高实体识别的效果。从各模型F1值可以看出，本文模型的整体识别效果得到了明显提升。

3 结语

针对实体增强预训练模型的词典获得方法较为复杂而且获取的实体词数量和使用范围有限的问题，本文充分利用了开放域知识库资源，使得词典的获得更加便利，能够包含更多相关领域的候选词，从而提升了模型的效果。在 NER 任务中，由于只依赖 BiLSTM 对标签打分的输出会导致出现大量不合法标签，本文通过加入 CRF 层的解码得到最优序列，提高了实体提取的结果。实验结果表明，利用加入知识库的预训练模型以及在 NER 任务中加入 CRF 解码层的模型获得了更高的 F1 值，从而验证了本文模型的有效性。未来的工作重点是简化模型，以提升模型的训练速度。

参考文献

[1] RIEDEL S, YAO LM, MCCALLUM A, et al. Relation extraction with matrix factorization and universal schemas[C]//Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2013: 74-84.

[2] CHEN YB, XU LC, ZENG DJ. Event extraction via dynamic multi-pooling convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2015: 167-176.

[3] DIEFENBACH D, LOPEZ V, SINGH K, et al. Core techniques of question answering systems over knowledge bases: a survey[J]. Knowledge and Information systems, 2018, 55(3): 529-569.

[4] 李源,马磊,邵党国,等.用于社交媒体的中文命名实体识别[J].中文信息学报,2020,34(8):61-69. (LI Y, MA L, SHAO DG, et al. Chinese Named Entity Recognition for Social Media[J]. Journal of Chinese Information Processing , 2020, 34(8): 61-69.)

[5] 张毅,王爽胜,何彬,等. 基于 BERT 的初等数学文本命名实体识别方法[J/OL]. 计算机应用:1-8[2021-07-08]. http://kns.cnki.net/kcms/detail/51.1307.TP.20210508.1604.008.html. (ZHANG Y, WANG SS, HE B, et al. Named entity recognition method of elementary mathematical text based on BERT[J/OL]. Journal of Computer Application: 1-8[2021-07-08].

[6] 李韧,李童,杨建喜,等. 基于 Transformer-BiLSTM-CRF 的桥梁检测领域命名实体识别[J]. 中文信息学报, 2021, 35(4): 83-91. (LI R, LI T, YANG JX, et al. Bridge Inspection Named Entity Recognition Based on Transformer-BiLSTM-CRF[J]. Journal of Chinese Information Processing, 2021, 35(4): 83-91.)

[7] LIU LY, SHANG JB, REN X, et al. Empower sequence labeling with task-aware neural language model[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018: 5253-5260.

[8] LI HB, MASATO H, LI Q, et al. Comparison of the Impact of Word Segmentation on Name Tagging for Chinese and Japanese[C]// Proceedings of the Ninth International Conference on Language Resources and Evaluation. Paris: European Language Resources Association, 2014: 2532-2536.

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

[10] MA RT, PENG ML, ZHANG Q, et al. Simplify the Usage of Lexicon in Chinese NER[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 5951–5960.

[11] LI XN, YAN H, QIU XP, et al. FLAT: Chinese NER Using Flat-Lattice Transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2020: 6836–6842.

[12] XUE MG, YU BW, LIU TW, et al. Porous lattice transformer Encoder for Chinese NER[C]// Proceedings of the 28th International Conference on Computational Linguistics (COLING). Stroudsburg, PA: Association for Computational Linguistics, 2020: 3831–3841.

[13] GUI T, MA RT, ZHANG Q, et al. CNN-Based Chinese NER with lexicon rethinking[C]// Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI). San Francisco: Morgan Kaufmann, 2019: 4982–4988.

[14] GUI T, ZOU YC, ZHANG Q, et al. A Lexicon-Based Graph Neural Network for Chinese NER[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA: Association for Computational Linguistics, 2019: 1040-1050.

[15] DEVLIN J, CHANG MW, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics, 2019:4171–4186.

[16] SUN Y, WANG SH, LI YK, et al. Ernie: Enhanced representation through knowledge integration [EB/OL]. (2019) [2020-01-21]. http://arxiv.org/abs/1904.09223.pdf.

[17] JIA C, SHI YF, YANG QR, et al. Entity Enhanced BERT Pre-training for Chinese NER[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA: Association for Computational Linguistics,2020: 6384—6396.

[18] XU B, XU Y, LIANG JQ, et al. CN-DBpedia: A Never-Ending Chinese Knowledge Extraction System[C]// Proceedings of the 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems. Cham: Springer, 2017: 428-438.

[19] XU L, TONG Y, DONG QQ, et al. CLUENER2020: Fine-grained named entity recognition dataset and benchmark for Chinese [EB/OL]. [2020-01-24]. https://arxiv.org/abs/2001.04351.

[20] LEVOW GA. The third international chinese language processing bakeoff: Word segmentation and named entity recognition[C]// Proceedings of the Fifth Workshop on Chinese Language Processing. Stroudsburg, PA: Association for Computational Linguistics, 2006: 108–117.

[21] BOUMA G. Normalized (pointwise) mutual information in collocation extraction[C] // Proceedings of the 2009 Conference on Biennial GSCL., Tübingen:Narr Franceke Attempto, 2009:31-40.

[22] SUN Y, WANG SH, LI YK, et al. Ernie 2.0: A continual pre-training framework for language understanding[C]//Proceedings of the 2020 Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2020 8968-8975.

This work is partially supported by the National Natural Science Foundation of China (61977021)， Guangzhou Key Laboratory of Big Data and Intelligent Education (201905010009).

HU Jie, born in 1977, Ph. D., professor. Her research interests include complex semantic big data management, natural language processing.

HU Yan, born in 1993, M.S. candidate. Her research interests include natural language processing.

LIU Mengchi, born in 1962, Ph.D., Professor. His research interests include semantic database, deep learning.

ZHANG Yan, born in 1974, Ph.D., Professor. His research interests include software engineering, information security.