计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900015

融合 BERT 和记忆网络的实体识别

陈 德 宋华珠 张 娟 周泓林
武汉理工大学计算机科学与技术学院 武汉 430070
(chende@whut.edu.cn)

摘 要 实体识别是信息提取的子任务,传统实体识别模型针对人员、组织、位置名称等类型的实体进行识别,而在现实世界中必须考虑更多类别的实体,需要细粒度的实体识别。同时,BiGRU 等传统实体识别模型无法充分利用更大范围内的全局特征。文中提出了一种基于命名记忆网络和 BERT 的实体识别模型,记忆网络模块能够记忆更大范围的特征,BERT 语言预训练模型能进行更好的语义表示。对水泥熟料生产语料数据进行实体识别,实验结果表明,所提方法能够识别实体且较其他传统模型更具优势。为了进一步验证所提模型的性能,在 CLUENER2020 数据集上进行实验,结果表明,在 BiGRU-CRF 模型的基础上使用 BERT 和记忆网络模块进行优化是能够提高实体识别效果的。

关键词:实体识别;BERT;记忆网络;BiGRU-CRF
中图法分类号 TP391

Entity Recognition Fusing BERT and Memory Networks

CHEN De,SONG Hua-zhu,ZHANG Juan and ZHOU Hong-lin
School of Computer Science and Technology,Wuhan University of Technology,Wuhan 430070,China

Abstract Entity recognition is a sub task of information extraction. The traditional entity recognition model is used to identify entities of personnel,organization,location and name. In the real world,more types of entities must be considered,and fine-grained entity recognition is needed. At the same time,traditional entity recognition models such as BiGRU cannot make full use of the global features in a wider range. This paper presents an entity recognition model based on memory network and BERT. The pre-training language model of BERT is used for better semantic representation,and the memory network module can memorize a wider range of features. The results of entity recognition for cement clinker production corpus data show that this method can recognize entities and has some advantages over other traditional models. In order to further verify the model in this paper,experiments are carried out on the CLUENER2020 dataset. The results show that the optimization based on BiGRU-CRF model using BERT and memory network module can improve the effect of entity recognition.

Keywords Entity recognition,BERT,Memory network,BiGRU-CRF

1 引言

开放领域的实体识别是一个热门的研究领域,其仅识别人名、地名、机构名称等实体,而比较火热的医疗领域的实体识别也仅识别疾病名称、症状、治疗方法等实体。但实际生产生活中的知识模式复杂,涉及到的实体类别多,需要进行细粒度的实体识别。而经典的双向门控循环网络(BiGRU)在充分利用全局信息方面存在局限性:在每一步的计算中,BiGRU 都将当前单词嵌入和过去的摘要状态用作输入,因而难以捕获句子级别的信息以及在整个数据源中的特征信息。针对这个问题,本文利用已有的知识信息来进行细粒度的实体识别,即在实体识别模型中结合 BERT 预训练语言模型。为了提取更大范围的特征,可使用记忆网络模块,其能够在整个数据集级别提取语义特征。最后,本文提出了一种基于 BERT 和记忆网络的实体识别方法,实体识别可以看作是序列标记任务。BERT 能够结合外部信息更好地进行语义表示,记忆网络模块能够编码更大范围的语义信息,BiGRU 对输入句子词嵌入向量序列进行语义编码,CRF 层将 BiGRU 层的输出和记忆网络模块的输出作为输入,可以对句子序列特征编码进行解码,最后输出实体识别结果,即在 BiGRU-CRF 的基础上,将 BERT 和记忆网络相融合来进行实体识别。

2 相关工作

实体识别技术主要分为基于传统机器学习的方法和基于

到稿日期:2020-09-02 返修日期:2020-12-06
基金项目:国家科技部科技基础性工作专项(2014FY110900)
This work was supported by the National Special Scientific and Technological Basic Work of the Ministry of Science and Technology (2014FY110900).
通信作者:宋华珠(shuaz@whut.edu.cn)

深度学习的方法。基于传统统计机器学习的方法主要包括词频逆词频(Term Frequency-Inverse Document Frequency, TF-IDF)[1]、隐马尔可夫(Hidden Markov Mode, HMM)[2]、最大熵(Maximum Entropy, ME)[3]、条件随机场(Conditional Random Fields, CRF)[4]以及支持向量机(Support Vector Machine, SVM)[5]。基于深度学习的方法主要包括基于循环神经网络(Recurrent Neural Network, RNN)[6]、长短期记忆网络(Long Short-Term Memory, LSTM)[7]和门控制循环神经网络(Gated Recurrent Unit, GRU)[8]等,它们都是通过提取实体上下文的语言特征以识别实体。

国内外学者对实体识别进行了许多研究。预训练语言模型指一串词序列的概率分布,是某种特定语言本身的特定描述。在传统实体识别模型中加入预训练语言模型能够提高词的语义表示,提升实体识别的效果。Peters 使用 BiLSTM 网络在大量语料上提取语言特征,使预训练的词表示能够包含丰富的句法和语义信息[9]。Devlin 提出使用语义能力更强的 Transformer 网络结构,同时使用 self-attention 机制使模型上下层直接全部互相连接,通过大规模语料预训练后,提取语义信息对数据进行语义编码,作为 NLP 模型的输入,其对各种 NLP 任务都能取得不错的效果[10]。Huang 等在双向长短期记忆网络中加入了条件随机场,提出了经典的 BiLSTM-CRF 模型,使网络能够学习句子级别的语意特征以识别实体[11]。Strubell 等提出采用空洞卷积神经网络(IDCNN-CRF)进行实体识别,在提取序列信息的同时加快了训练速度[12]。Ling 等将细粒度的实体识别任务视作一个多标签分类任务,提出了 FIGER 模型,并实现了一个 FG-NER 系统以准确地预测实体类型标签[13]。Mai 等对常见的细粒度的实体识别方法进行了总结,并在英语和日语语料数据集上进行实验,针对在英语数据集上识别效果最好但在日语语料数据集上识别效果差的问题,提出利用词典和实体类别标签嵌入信息对 LSTM+CNN+CRF 模型进行优化[14]。

3 融合 BERT 和记忆网络的实体识别模型

本文的实体识别模型的总体结构如图 1 所示。嵌入层将输入转换成计算机能够计算的向量形式,同时保留了一定语义表示。嵌入层基于有 BERT 预训练语言向量的向量表示,融合外部语言本身的知识,经过嵌入层对每个字符和外部知识的向量表示,得到输入字符向量的序列。BiGRU 对其进行进一步的语义编码,将 BiGRU 层输出作为关键字查询记忆网络组件以获取数据集级特征,再将记忆网络模块的输出和 BiGRU 层的输出作为 CRF 层的输入。最后,通过 CRF 层输出标记序列的最大概率。

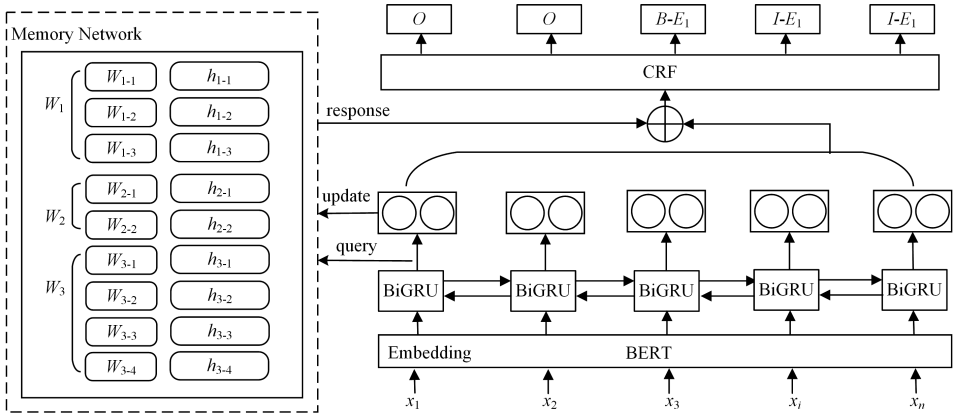

图 1 融合 BERT 和记忆网络的实体识别模型

Fig. 1 Entity recognition model combining BERT and memory network

3.1 BERT 模块

通过大量预训练深度神经网络模型,可以提取语言本身的特征[15]。在此基础上,当在语料数据中进行模型训练时可以进行参数微调,从而在实体识别任务的场景中可以更好地对数据进行语义表示。常见的语言模型方法是从左到右计算下一个词的概率,如式(1)所示:

$$p(S)=p(w_1,w_2,\cdots,w_m)=\prod_{i=1}^{m}p(w_i\mid w_1,w_2,\cdots,w_{i-1}) \tag{1}$$

但在具体的实体识别任务中,需要的不是一个完整的语言模型,而是需要一个字的上下文表示,能够表达字的语义。

BERT(Bidirection Encoder Representations from Transformers)的模型结构如图 2 所示。为了提取字上下文的特征,BERT 采用双向 Transformer 作为语义编码器。

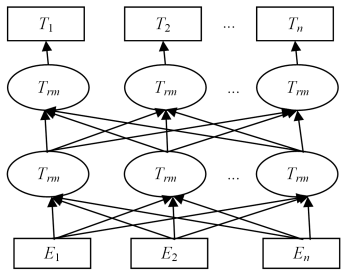

图 2 BERT 预训练语言模型

Fig. 2 BERT pre-training language model

BERT 模型基于“Masked”和“下一个句子预测”两个任务来训练模型,分别提取词级别和句子级别的特征。“Masked”任务是为了训练深度双向语言表示向量,通过遮住句子中的某些词,再让编码器预测这个单词的原始词汇。“下一个句子预测”任务是训练一个二分类的模型,句子间的语言

关系能反映语言的特征，通过提取句子之间的语义关系来训练模型。即这个任务在每个训练前的例子中选择句子 A 和句子 B，50%的训练数据中 B 是在 A 后面的下一个句子；50%的训练数据中 B 不是在 A 后面的下一个句子，而是训练数据中其他的随机句子。通过提取这两个句子的语义特征来预测是否是下一句。这种训练能使模型更容易地提取到语料数据的语义表示。

BERT 模型最关键的组件是 Transformer 编码结构，其采用基于注意力的机制对文本进行编码而不是采用传统的 RNN 网络。Transformer 编码结构单元如图 3 所示。编码单元中核心模块是自注意力部分，如式(2)所示：

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

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 均是输入字向量矩阵，d_k 是输入向量维度。计算句子中的每个词对于这句话中所有词的相互关系，然后这些词与词之间的相互关系在一定程度上反映了这个句子中不同词之间的关联性以及重要程度。利用这些相互关系来调整每个词的重要性(权重)就可以获得每个词新的语义特征的表达。这个新的表征不但蕴含了该词本身，还蕴含了其他词与这个词的关系，因此与单纯的词向量相比，Transformer 语义编码更具有全局的语义表示信息。

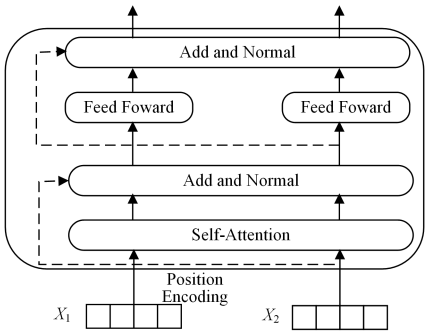

图 3　Transformer 编码单元

Fig. 3　Transformer encoding unit

为了扩展模型专注于不同位置的表达能力，增大注意力单元的表示子空间，Transformer 采用“多头”模式，如式(3)、式(4)所示：

$$MultHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(head_1,\cdots,head_h)W^o \tag{3}$$

$$head_i=Attention(\boldsymbol{Q}W_i^Q,\boldsymbol{K}W_i^k,\boldsymbol{V}W_i^V) \tag{4}$$

对于实体识别任务而言，语料数据中的词在句子中的位置特征非常重要，由于一般的自注意力机制无法抽取时序特征，因此 Transformer 模块采用了位置嵌入的方式来加入位置特征，如式(5)、式(6)所示：

$$PE(pos,2i)=\sin(pos/10\,000^{2i/d_{\mathrm{model}}}) \tag{5}$$

$$PE(pos,2i+1)=\cos(pos/10\,000^{2i/d_{\mathrm{model}}}) \tag{6}$$

3.2　BiGRU 文本特征编码模块

门控制循环单元是由长短记忆神经网络简化而来的，即将遗忘门和输入门合并成一个更新门，相对于 LSTM 而言，其结构更简单，参数更少，训练更加高效。其设计的特点非常适合对时序数据进行建模，如文本数据。GRU 模型结构如图 4 所示，x_t 代表 t 时刻的输入，z_t 代表更新门，r_t 是一个重置门，用于控制信息丢失，h_t 是隐层状态，其中 x_t 包含 BERT 字嵌入和对应的词特征嵌入。

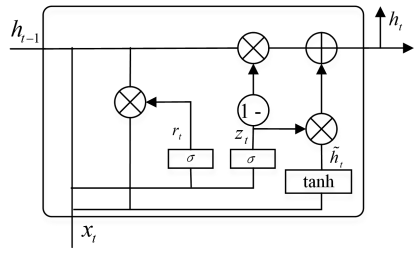

图 4　GRU 编码单元结构

Fig. 4　GRU coding unit structure

更新门用于控制前一时刻的状态信息被带入到当前状态中的程度，更新门的值越大，说明前一时刻的状态信息被带入得越多。重置门用于控制忽略前一时刻的状态信息的程度，重置门的值越小，说明忽略得越多。在 GRU 的计算过程中，通常用式(7)－式(10)来计算各个门控制的信息。

$$z_t=\sigma(\boldsymbol{W}_i*[h_{t-1},x_t]) \tag{7}$$

$$r_t=\sigma(\boldsymbol{W}_r*[h_{t-1},x_t]) \tag{8}$$

$$h_t=\tanh(\boldsymbol{W}_c*[r_t\cdot h_{t-1},x_t]) \tag{9}$$

$$h_t=(1-z_t)\cdot c_{t-1}+z_t\cdot h_t \tag{10}$$

其中，x_t 指当前时刻的输入，σ 通常是向量相加或者相乘的 sigmoid 函数，$\cdot$ 是两个向量对应的点积，$\boldsymbol{W}_i$，$\boldsymbol{W}_r$ 和 $\boldsymbol{W}_c$ 代表权重矩阵，$*$ 表示矩阵的乘积。

标准 GRU 按文本序列接收输入，其只能处理前文信息而忽略了下文信息。双向门控制循环单元网络(BiGRU)对每一个输入序列包含一个 forward 和 backward 的 GRU 网络，BiGRU 网络的输出结果由这两个 GRU 网络共同作用得到，可以对输入句子序列提取双向语义特征信息。即 BiGRU 结合正向 GRU 网络和逆向 GRU 网络，得到了更好的效果。

$$h_t=\overrightarrow{h}_t\times\overleftarrow{h}_t \tag{11}$$

3.3　记忆网络编码模块

一般实体识别模型仅仅从词、字符的前一个或者后一个词、字符来提取上下文特征，也有基于 CRF 提取句子级的语义特征。而本文基于记忆网络模块来记忆数据在整个数据源中的上下文环境的语言特征，并在 BERT 和 BiGRU-CRF 模型的基础上融合记忆网络[16]模块来识别数据语料中的实体。例如，在对输入 x_i 编码时会从记忆网络模块中取出保存的特征 r_i，然后与 BiGRU 层的输出 h_i 融合作为 CRF 层的输入，同时更新记忆网络模块的记忆特征值。基于前文 BERT-BiGRU-CRF 模型使用了记忆网络组件，对整个数据源中出现过的实体的特征编码进行“记忆”可以从数据源层的语义特征进行融合，数据源级别的语义特征使用 key-value 记忆网络组件来保持上下文的语义特征。

采用键值记忆网络组件 M 来存储文档级别的上下文表示。内存插槽定义为向量对$(\boldsymbol{k}_1,\boldsymbol{v}_1),\cdots,(\boldsymbol{k}_m,\boldsymbol{v}_m)$。在每个单个插槽中，键表示词 w_i 嵌入，其值是训练实例中每个令牌的

序列标记编码器 h_i 对应的隐藏状态。由于在不同上下文中更改嵌入和表示，同一单词可能会占据许多不同的位置，因此使用训练实例来帮助指示所查询令牌的网元类型的示例。

(1)记忆更新

单词嵌入在训练过程中经过了微调，可用于更新记忆的关键部分。序列标签编码器的隐藏层状态改变，记忆值也将随之改变。假设第 i 个令牌的状态在计算后发生了变化，则存储器 M 中的第 i 个插槽将被重写，每个记忆插槽将更新一次。

(2)记忆查询

对于句子中的第 i 个单词，我们通过反向索引找到该单词在内存 M 中的所有上下文表示，该反向索引找到一个大小为 T 的子集 $(k_{sub1}, v_{sub1}), \cdots, (k_{subT}, v_{subT})$，其中反向索引记录唯一一个单词在存储器 M 中的位置。T 表示训练数据中该单词出现的次数。对于一个词，将存储键 $k_j \in [k_{sub1}; \cdots; k_{subT}]$ 用作关注键，将存储值 $v_j \in [v_{sub1}; \cdots; v_{subT}]$ 用作注意力分数，然后将查询词的嵌入 w_{qi} 用作注意力查询 q_i。用 $u_{i,j}$ 表示 q_i 和 k_j 的关联性，考虑点乘、缩放点乘和余弦相似度 3 种方式来计算关联性，如式(12)所示：

$$u_{i,j}=\begin{cases} q_i k_j^{\mathrm{T}} \\ \dfrac{q_i k_j^{\mathrm{T}}}{\sqrt{d_w}} \\ \dfrac{q_i k_j^{\mathrm{T}}}{\| q_i \| \| k_j \|} \end{cases} \tag{12}$$

其中，d_w 表示词嵌入向量的维数。

(3)记忆组件查询返回

数据源级的语义表示通过式(13)和式(14)计算得到，$\alpha_{i,j}$ 表示水泥熟料生产记忆网络组件模块返回查询 q_i 和对应的 k_j 在所有返回查询中所占的比例，而 r_i 表示这些返回查询的综合值，其值等于所有 $\alpha_{i,j}$ 和对应的 v_j 值的加权和。

$$\alpha_{i,j}=\frac{\exp(u_{i,j})}{\sum_{z=1}^{T}\exp(u_{i,z})} \tag{13}$$

$$r_i=\sum_{j=1}^{T}\alpha_{i,j}v_j \tag{14}$$

其中，v_j 表示记忆网络组件插槽中的 q_i 对应的值。原隐藏层状态的融合表示 $g_i \in \mathbb{R}^{d_h}$ 和文档层的语义表示作为最后 CRF 层的输入，g_i 的计算方法如式(15)所示：

$$g_i=\lambda h_i+(1-\lambda)r_i \tag{15}$$

其中，d_h 表示隐藏层维数的大小；λ 是超参数，表示数据源级的语义信息对识别信息的贡献，当 $\lambda=1$ 时则表示不考虑数据源级的语义信息。

3.4 序列标注模块 CRF

利用 CRF 模型则不会出现这种错误，因为 CRF 的特征函数的存在就是为了对输入序列进行观察、学习，得到其特征。因此，考虑前后标签约束关系，加入 CRF 层，将 BiGRU 与 CRF 进行结合，这样既可以利用 BiGRU 提取文本序列中的上下文信息，也可以通过 CRF 在整句层面上的标注信息来提高标注精确率，从而获取全局最优输出序列，得到更好的实体识别效果。CRF 层将 BiGRU 层的输出和记忆网络的查询返回矩阵 $\boldsymbol{P}$ 作为输入。在 CRF 层计算一个转移矩阵 $\boldsymbol{A}$，则 $A_{i,j}$ 代表输入的第 i 个字符的语义编码到标签 j 之间的转移概率。在给定输入序列 x 的条件下输出标签序列 y 的概率分数如式(16)所示：

$$s(x,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=0}^{n}P_{i,y_i} \tag{16}$$

整个序列的分数等于各个位置的分数之和，根据输入序列 x 对应的输出标签序列 y 计算出的这个分数，选择出分数最大的一个作为最终的输出标签序列。最后，利用 softmax 函数进行归一化处理，如式(17)所示：

$$P(y|x)=\frac{\exp(s(x,y))}{\sum_{y'}\exp(s(x,y))} \tag{17}$$

4 实验

(1)实体识别实验设置

实体识别实验的硬件环境如下：CPU 为 E5 2680 v2，内存为 64 GB，GPU 为 Tesla T4，其内存为 16 GB，硬盘为 SSD，其容量为 512 GB。实验软件环境如下：操作系统为 Ubuntu 16.04，Python 3.7，Tensorflow-GPU 1.15.0，Jieba 0.39。本文的实体识别模型使用的损失函数是均方差函数[14]，其他参数如下。

embedding_dim=64	#词向量维度
seq_length=200	#序列长度
num_classes=10	#类别数
vocab_size=5000	#词汇表大小
hidden_dim=128	#双向 GRU 层神经元
dropout_keep_prob=0.5	# dropout 保留比例
learning_rate=1×10^{-3}	# 学习率
batch_size=64	#每批训练大小
num_epochs=15	#总迭代轮次

从水泥企业网站和知网收集了水泥熟料生产领域的数据，最终爬取了 2611 个知网页面和 1321 个水泥企业网站页面，并基于术语词典的数据标注总共标注了 11096 个句子，包含 12927 个实体。将实验数据中的实体按照其类别进行统计，水泥熟料生产领域数据顶层十大类别的实体分布统计如图 5 所示。

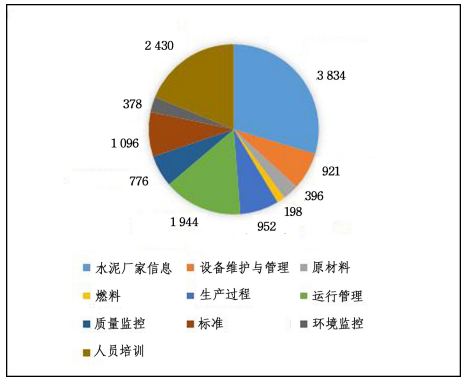

图 5 水泥熟料生产领域数据顶层实体分布统计(电子版为彩色)

Fig. 5 Top-level entity distribution statistics in the field of cement clinker production data

图 6 给出了水泥熟料生产领域数据中更细分的实体类别的分布统计。按照 4∶1 的比例将标注的语料数据分为训练集和测试集，即随机选取 8877 个句子作为训练集，2219 个句子作为测试集。

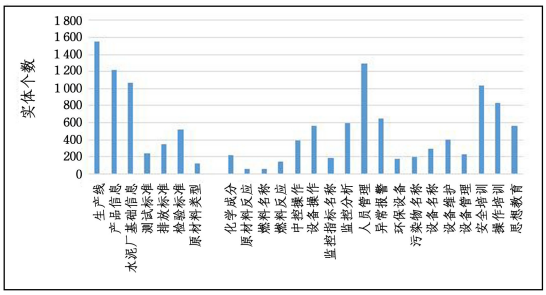

图 6　水泥熟料生产领域数据实体分布统计

Fig. 6　Data entity distribution statistics in the field of cement clinker production

由于水泥熟料生产领域没有公开的标注好的 benchmark 数据集，为了进一步验证本文提出的基于 BERT 和记忆网络的实体识别模型的性能，在公开数据集 CLUENER2020 上进行了实验验证。CLUENER2020 数据集是在文本分类数据集 THUCTC 的基础上，选取其中部分数据对 10 种实体类别命名实体进行标注。CLUENER2020 数据集中，有 10748 个句子用于训练集和 1343 个句子用于测试集，总共包含 10 种实体类别：地址、书名、公司、游戏、政府、电影、姓名、组织机构、职位和景点。各类实体数量分布如图 7 所示。

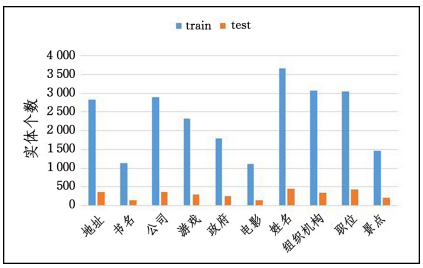

图 7　CLUENER2020 数据统计

Fig. 7　CLUENER2020 data statistics

(2)水泥熟料生产的实体识别实验

为了验证本文模型在水泥熟料生产领域的实体识别性能，选择经典的实体识别模型 Dictionary Match，CRF，BiLSTM，BiLSTM-CRF 和 BiGRU-CRF 与本文的实体识别模型进行对比。用记忆网络组件来提取语义特征，结合 BiGRU 层的输出作为 CRF 层的输入。利用深度学习模型开发框架 Keras 来搭建本文模型，简化网络模型的实现。本文研究的水泥熟料生产领域没有公开的标注好的 benchmark 数据集，基于术语词典标注本身存在标注错误以及术语实体词典能覆盖所有实体，存在未包含在术语词典中的实体情况。鉴于以上问题，为了评估实体识别的效果，只能随机抽取部分最终模型识别出的实体来进行人工判断。与其他模型的对比实验使用精确率、召回率和 F1 值作为评价指标，随机抽取 50 个句子，通过人工判断实体识别方法来识别正确的实体、错误的实体以及没有识别出的实体。将 Dictionary Match，CRF，BiLSTM，BiLSTM-CRF，BiGRU-CRF 实体识别模型以及本文方法进行对比，结果如表 1 所列。

表 1　不同实体识别模型的实验对比结果

Table 1　Experimental comparison results of different entity recognition models

（单位：%）

方法	Precision	Recall	F1
Dictionary Match	82.16	23.84	36.95
CRF	46.65	41.33	43.82
BiLSTM	53.98	49.74	51.77
BiLSTM-CRF	58.39	60.15	59.25
BiGRU-CRF	62.82	56.36	59.41
本文方法	73.85	69.35	71.53

基于词典匹配的 Dictionary Match 方法的精确率较高而召回率较低，目前词典不能完全覆盖数据中出现的实体，对于词典中没有出现的实体效果非常差。而 CRF 模型基于统计的方法，其精确率、召回率以及 F1 值分别为 46.65%，41.33%和 43.82%。CRF 模型的精确率没有 Dictionary Match 方法的高，但是召回率和 F1 值优于 Dictionary Match。从整体上看，BiLSTM 模型、BiGRU-CRF 模型和本文方法等基于深度学习的实体识别方法的识别效果优于 Dictionary Match 和 CRF 传统方法，一方面深度学习模型复杂且模型参数也多，另一方面深度学习模型需要进行更多的训练和计算。BiLSTM-CRF 模型的精确率、召回率以及 F1 值分别为 58.39%，60.15%，59.25%，比 BiLSTM 模型分别提高了 4.41%，10.41%，7.48%。深度神经网络加入 CRF 后能在句子层范围提取更好的序列特征，显著提高了模型的性能。而 BiGRU-CRF 模型在识别效果上仅略微优于 BiLSTM-CRF 模型，这是由于两者模型的结构比较相似。本文的实体识别方法是针对水泥熟料生产的实体识别问题，并基于 BiGRU-CRF 模型进行优化的，精确率、召回率以及 F1 值分别为 73.85%，69.35%和 71.53%。从实验结果来看，本文模型优于其他对比模型，本文的实体识别方法具有一定优势。

(3)CLUENER2020 实体识别实验及结果分析

为了进一步验证本文的实体识别模型的有效性，在标准数据集 CLUENER2020 上进行实验，实验初始参数如前文水泥数据的实体识别所示。同时，与 BERT，BiGRU-CRF 及 BERT-BiGRU-CRF 进行对比实验，以验证本文算法的性能。

本文的对比实验中 BERT 模型的初始参数设置如下：embedding 维数为 64，batch_size 为 32，最大序列长度为 64，学习率为 3×10^{-5}，词嵌入词汇数量大小为 21128，隐层维度为 768，隐层层数为 12，注意力头为 12，优化器为 Adam，损失函数为均方差函数，训练迭代的 epochs 次数为 15；BiGRU-CRF 模型的初始参数设置如下：embedding 维数为 64，batch_size 为 32，隐层维度为 384，隐层层数为 2，优化器为 Adam，损失函数为均方差函数，训练迭代的 epochs 次数为 15。

本文模型使用的损失函数是均方差函数。由图 8 和图 9 可以看出，在训练过程中，本文模型的 loss 随着模型的训练一直下降，而精确率在训练过程开始时最低，然后逐渐升高，中间有所波动，当 epoch 到第 10 次后精确率的值趋于稳定。

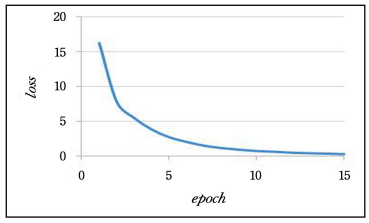

图 8 本文模型的 loss 在训练过程中的变化曲线

Fig. 8 Loss curve of the model in this paper during thetraining process

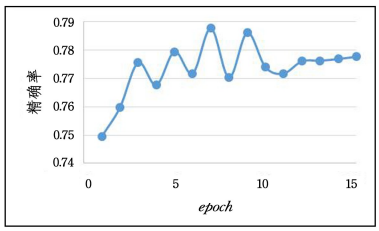

图 9 训练过程中本文模型的精确率的收敛情况

Fig. 9 Convergence of the accuracy of the model in the training process

本文基于 BERT 和记忆网络的实体识别模型在数据集 CLUENER2020 上的实验结果表 2 所列。在所有十大类实体中，姓名类别实体的实验结果最好，精确率、召回率和 F1 值分别为 86.71%，88.39%，87.54%，姓名类别的实体个数在数据集中也是最多的，模型相对而言学习到更多的姓名类实体的特征，相对识别效果较好。而地址和景点类型的实体识别效果最差，地址的精确率、召回率和 F1 值分别为 61.56%，63.54%，62.53%，而景点的分别为 68.9%，68.9%，68.9%。地址和景点类型的实体含义相似，如景点中有“故宫”，而“故宫”也可以看作一个地址；地址类型的实体如“中国电竞馆”也可看作一个景点。这个分类不够准确，而且现在的数据集的数据量相对而言还不能充分挖掘这两种类别实体的特征信息，因此这两种类别实体的识别效果较差。而其他类型的命名实体的识别效果都较好，识别的精确率、召回率和 F1 值基本都接近 80% 或更高。

表 2 本文方法在数据集 CLUENER2020 上的实验结果

Table 2 Experimental results of the method in the dataset CLUENER2020

（单位：%）

实体	Precision	Recall	F1
姓名	86.71	88.39	87.54
组织机构	78.02	79.29	78.65
职位	80.75	79.45	80.09
公司	80.66	83.86	82.23
地址	61.56	63.54	62.53
游戏	81.9	87.46	84.59
政府	79.03	85.43	82.1
景点	68.9	68.9	68.9
书名	86.33	77.92	81.91
电影	80	82.12	81.05
平均值	77.77	83.25	80.42

为了进一步验证本文方法的有效性，本文进行了对比实验以验证优化的模块对实体识别效果的提升。使用 BERT，BiGRU-CRF，BERT-BiGRU-CRF 和本文提出的基于记忆网络模块的 BERT-BiGRU-CRF-Mem 模型，采用 F1 值作为评价指标，在 CLUENER2020 数据集上对不同实体进行识别，实验结果如表 3 所列。本文基于 BERT 和记忆网络的实体识别模型的识别效果较好，总体的 F1 值为 80.42%。

表 3 CLUENER2020 数据集上 F1 值的对比实验结果

Table 3 Comparative experimental results of F1 value on CLUENER2020 dataset

（单位：%）

实体	BERT	BiGRU-CRF	BERT-BiGRU-CRF	本文方法
姓名	77.42	74.04	88.75	87.54
组织机构	54.94	75.96	79.43	78.65
职位	70.39	70.16	78.89	80.09
公司	69.07	72.27	81.42	82.23
地址	50.82	45.5	60.89	62.53
游戏	73.38	85.27	86.42	84.59
政府	74.28	77.25	87.03	82.1
景点	58.18	52.42	65.1	68.9
书名	73.33	67.2	73.68	81.91
电影	63.33	78.97	85.82	81.05
平均值	66.51	70	78.82	80.42

同样地，所有模型在地址和景点这两类命名实体上的识别效果均较差。利用单纯的 BERT 模型来进行实体识别的效果较差，其 F1 值在所有对比实验模型中也最低，为 66.51%。BERT 模型可以对输入字符进行很好的语义表示，但对于输入句子中字符间的语义关系的获取较差。而常规实体识别模型 BiGRU-CRF 的 F1 值略高于 BERT 模型，分别为 70%和 66.5%。利用双向的 GRU 进行语义编码，将 CRF 层作为语义解码，能够对句子中字符前向和后向的相互关系提取特征，从而更好地获取整个句子的特征。而在此基础上采用预训练语言模型 BERT 来进行语义表示，然后将其作为 BIGRU 层的输入，BERT-BiGRU-CRF 模型识别的 F1 值为 78.82%，相对于单纯的 BiGRU-CRF 模型，BERT-BiGRU-CRF 模型的整体 F1 值提高了 8.82%。同样包含 BERT 模块的记忆网络 BERT-BiGRU-CRF 模型的整体识别的 F1 为 80.42%，BERT 预训练模型的加入能大幅度提升模型性能。基于记忆网络的 BERT-BiGRU-CRF-Mem 模型在所有实验模型中性能最好，模型中记忆网络模块对输入字符进行编码，提取数据源级的特征，模型训练过程中特征值动态更新，和 BiGRU 层的输出共同作为 CRF 层的输入。相比 BERT-BiGRU-CRF 模型，本文提出的基于记忆网络的 BERT-BiGRU-CRF-Mem 模型的 F1 提升了 1.6%。综上，本文提出的 BERT- BiGRU-CRF-Mem 模型中的 BERT 和记忆网络模块对实体识别有提升效果。

本文算法模型中包含 103 821 832 个参数，模型的训练时间为 2 h16 min；而 BiGRU-CRF 模型仅包含 602 552 个参数，模型的训练时间为 20 min；BERT 模型包括 101 360 640 个参数，模型的训练时间为 55 min。BERT 模块非常复杂且包含的参数非常多，使得模型的训练时间更长。本文方法在 BERT 和记忆网络的基础上对实体识别进行优化，参数数量更多时，会使模型变得更复杂，训练时间也会更长。

目前还未对本文模型进行实时性场景的应用，但是目前模型的训练时间较长，在未来将考虑对训练过程进行优化以及设计能够适应更强大的 GPU 集群的算法来加速训练过程。同时考虑文献[17-20]的方法，进一步利用中文语言的特点，在已有工作的基础上对本文方法进行优化。

综上，本文提出的基于 BERT 和记忆网络的实体识别模型能够提取语义特征来对实体进行识别。在基础的 BiGRU-CRF 模型上加入 BERT 和记忆网络模块能够提高实体识别的效果。

结束语　传统的命名实体识别模型只是将数据集中的句子作为输入，而未考虑数据集级别的特征。本文使用记忆网络模块对全局数据集级别的特征进行编码，并使用 BERT 获得更好的语义表示，提出了融合 BERT 和记忆网络的实体识别模型。将其应用到水泥熟料生产领域的实体识别中，实验结果表明，本文方法能够识别水泥熟料生产领域的实体且较传统方法具有一定优势。为了进一步验证本文方法在 CLUENER2020 数据集上的优势，本文进行了对比实验，结果表明，与 BERT，BiLSTM-CRF，BiGRU-CRF，BERT-BiGRU-CRF 深层神经网络相比，融合 BERT 和记忆网络可以提高实体识别的性能，获得更好的实体识别效果。在将来的工作中，我们将进一步探索可以提取更多全局特征的其他组件，并在命名实体识别任务中考虑实体之间的语义关系。

参 考 文 献

[1] GAO B T，ZHANG Y，LIU B. BioTrHMM：Biomedical named entity recognition algorithm based on transfer learning[J]. Application Research of Computers，2019，36(1)：45-48.

[2] YU X，ZHANG J，QIU W S，et al. Research on medical literature risk event extraction based on sequence annotation algorithm comparison[J]. Computer Applications and Software，2017，34：12.

[3] ZHOU X J，XU C M，RUAN T. Multi-granularity Medical Entity Recognition for Chinese Electronic Medical Records[J]. Computer Science，2021，48(4)：237-242.

[4] MCCALLUM A，LI W. Early results for named entity recognition with conditional random fields，feature induction and web-enhanced lexicons[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL. Stroudsburg：Association for Computational Linguistics，2003(4)：188-191.

[5] ISOZAKI H，KAZAWA H. Efficient support vector classifiers for named entity recognition[C]// Proceedings of the 19th International Conference on Computational Linguistics. Stroudsburg：Association for Computational Linguistics，2002(1)：1-7.

[6] LI S，LI W，COOK C，et al. Independently recurrent neural network (indrnn)：Building a longer and deeper rnn[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：5457-5466.

[7] TAI K S，SOCHER R，MANNING C D. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks[C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1：Long Papers). 2015：1556-1566.

[8] CHUNG J，GULCEHRE C，CHO K H，et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [EB/OL]. https：//arxiv. org/pdf/1412. 3555. pdf，2018.

[9] PETERS M E，NEUMANN M，IYYER M，et al. Deep contextualized word representations [C]// Proceedings of NAACL-HLT. 2018：2227-2237.

[10] DEVLIN J，CHANG M W，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，Volume 1 (Long and Short Papers). 2019：4171-4186.

[11] HUANG Z，XU W，YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. https：//arxiv. org/pdf/1508. 01991. pdf，2015.

[12] STRUBELL E，VERGA P，BELANGER D，et al. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions [C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017：2670-2680.

[13] LING X，WELD D S. Fine-grained entity recognition[C]// Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence. 2012：94-100.

[14] MAI K，PHAM T H，NGUYEN M T，et al. An empirical study on fine-grained named entity recognition[C]// Proceedings of the 27th International Conference on Computational Linguistics. 2018：711-722.

[15] LIU X，CHEN Q，DENG C，et al. Lcqmc：A large-scale chinese question matching corpus[C]// Proceedings of the 27th International Conference on Computational Linguistics. 2018：1952-1962.

[16] MILLER A，FISCH A，DODGE J，et al. Key-Value Memory Networks for Directly Reading Documents[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016：1400-1409.

[17] ZHANG H N，WU D Y，LIU Y，et al. Chinese named entity recognition based on deep neural network[J]. Journal of Chinese Information Processing，2017，31(4)：28-35.

[18] DONG Z，SHAO R Q，CHEN Y L，et al. Named Entity Recognition in Food Field Based on BERT and Adversarial Training[J]. Computer Science，2021，48(5)：247-253.

[19] ZHANG D，CHEN W L. Chinese Named Entity Recognition Based on Contextualized Char Embeddings [J]. Computer Science，2021，48(3)：233-238.

[20] ZHANG D，WANG M T，CHEN W L. Named Entity Recognition Combining Wubi Glyphs with Contextualized Character Embeddings[J]. Computer Engineering，2021，47(3)：94-101.

CHEN De，born in 1992，postgraduate. His main research interest includes natural language processing and so on.

SONG Hua-zhu，born in 1970，Ph.D，associate professor，master supervisor，is a senior member of China Computer Federation. Her main research interests include artificial intelligent and data mining，semantic and knowledge abstraction.