doi:10.3969/j.issn.1001-4616.2022.03.016

基于集成学习的中文命名实体识别方法

梁兵涛[1],倪云峰[2]

(1.杭州优行科技有限公司,浙江 杭州 310000)
(2.西安科技大学通信与信息工程学院,陕西 西安 710600)

[摘要] 针对中文命名实体识别经典的 BiLSTM-CRF(bi-directional long short-term memory-conditional random field)模型存在的嵌入向量无法表征多义词、编码层建模时注意力分散以及缺少对局部空间特征捕获的问题,本文提出一种融合 BERT-BiGRU-MHA-CRF 和 BERT-IDCNN-CRF 模型优势的集成模型完成命名实体识别. 该方法利用裁剪的 BERT 模型得到包含上下文信息的语义向量;再将语义向量输入 BiGRU-MHA(bi-directional gated recurrent unit-multi head attention)及 IDCNN(Iterated Dilated Convolutional Neural Network)网络. 前者捕获输入序列的时序特征并能够根据字符重要性分配权值,后者主要捕获输入的空间特征,利用平均集成方式将捕获到的特征融合;最后通过 CRF 层获得全局最优的标注序列. 集成模型在人民日报和微软亚洲研究院(Microsoft research asia,MSRA)数据集上的 $F1$ 值分别达到了 96.09% 和 95.01%. 相较于单个模型分别提高了 0.74% 和 0.55%以上,验证了本文方法的有效性.

[关键词] 命名实体识别,BERT 模型,集成学习,注意力机制,迭代膨胀卷积网络
[中图分类号]TP391　[文献标志码]A　[文章编号]1001-4616(2022)03-0123-09

Chinese Named Entity Recognition Method Based on Ensemble Learning

Liang Bingtao[1],Ni Yunfeng[2]

(1.Hangzhou Youxing Technology CO.,LTD.,Zhejiang 310000,China)
(2.College of Communication and Information Engineering,Xi'an University of Science and Technology,Xi'an 710600,China)

Abstract:Aiming at the problems existing in the classical BiLSTM-CRF(bi-directional long short-term memory-conditional random field) model of Chinese named entity recognition,such as the inability of the embedding vector cannot represent polysemy,the attention of the coding layer is distracted and lack of capturing local spatial features. This paper proposes an ensemble model that combines the advantages of the BERT-BiGRU-MHA-CRF and BERT-IDCNN-CRF models to complete named entity recognition. This method uses the BERT model to obtain a semantic vector containing contextual information,and then inputs the semantic vector into BiGRU-MHA(bi-directional gated recurrent unit-multi head attention) and IDCNN(Iterated Dilated Convolutional Neural Network) networks. The former captures the timing characteristics of the input sequence and can assign weights according to the importance of the characters,the latter mainly captures the spatial characteristics of the input,and uses the mean ensemble method to fuse the captured features. Finally,the global optimal annotation sequence is obtained through the CRF layer. The $F1$ values of the ensemble model on the datasets of People's Daily and Microsoft Research Asia(MSRA) reached 96.09% and 95.01%,respectively. Compared with the single model,it has increased by more than 0.74% and 0.55%,respectively,which verifies the effectiveness of the method in this paper.

Key words:named entity recognition,BERT model,ensemble learning,attention mechanism,IDCNN

命名实体识别(named entity recognition,NER)是自然语言处理领域的一项基础任务,主要目的是识别出文本中包含的如人名、地名和组织机构名等具有实际含义的实体. 随着算力和识别要求的提高,完成 NER 任务的方法也从基于字典和基于统计的传统方法转向结合注意力机制和迁移学习的基于深度学习的方法.

基于深度学习的命名实体识别按照所采用的神经网络类型可以分为基于循环神经网络(recurrent

收稿日期:2022-01-10.
通讯作者:倪云峰,教授,博士,研究方向:煤矿信息化技术研究等. E-mail:635129613@qq.com

neural networks,RNN)的命名实体识别和基于卷积神经网络(convolutional neural networks,CNN)的命名实体识别. RNN 网络因为其结构特性更适合对文本数据进行建模,因而在命名实体识别领域大多采用 RNN 结构的模型. 文献[1]首次在命名实体识别中利用长短时记忆网络解决了序列的长期依赖问题,而后 Lample 等人又在此基础上提出了 BiLSTM-CRF 模型[2],利用双向的长短期记忆网络同时捕获字词的上下文信息,使得模型具有更好的提取特征能力. 该模型在 CoNLL-2003 语料库上得到了 90.94%的 $F1$ 值,也使得该模型成为命名实体识别的经典模型. 由于中文文本没有明显的实体边界,先分词再标注的方式会产生错误传播问题,而文献[3]提出了基于字符的 Lattice LSTM 模型,利用栅格式 LSTM 网络从输入获得最相关的词信息,避免了分词错误,在 MSRA 数据集上达到了 93.18%的 $F1$ 值.

基于 RNN 的命名实体识别虽然适合对文本序列进行建模,捕获其时序特征,但其缺少对空间特征的建模. 如果以“我来自西安科技大学”为例进行实体识别,采用基于字符的方式进行建模,得到“我、…、西、安、科、技、大、学”,如果有一条特征能够同时提取到“西、安、科、技、大、学”这几个字组成的一个局部空间,则会大大提高模型将其识别为组织机构类实体的概率.

卷积神经网络能够并行地对输入的空间特征建模. 在命名实体识别任务中,也有研究者利用 CNN 网络对空间特征进行抽取. Gui 等提出了一种基于卷积神经网络的方法[4],该方法使用重新思考机制并融合了词典信息,并能够通过反馈高级特征来细化网络,解决单词冲突的问题. Wang 等提出一种基于门控机制的 CNN 模型[5],在 MSRA 数据集上得到了 91.23%的 $F1$ 值. 在使用 CNN 进行文本建模时,最后一层神经元只能获得输入文本中的一小段信息,而 NER 任务往往存在长期依赖问题,为了覆盖整个输入序列同时避免过拟合,往往需要叠加更多的卷积层和 Dropout 层,最终导致模型参数过多难以训练. Yu 等提出的膨胀卷积神经网络在卷积过程中加入了膨胀宽度这一参数[6],有效地解决了这一问题. Yu 等在此基础上利用迭代膨胀卷积神经网络提出一种武器装备领域的实体识别模型[7],并在测试集上获得了超过 94%的 $F1$ 值.

综上所述,RNN 适合对文本序列建模,但却未对文本中的重要词和普通词进行区分,且缺少对空间特征的考虑;CNN 适合对空间特征建模,但难以提取文本所包含的时序信息. 本文针对 BiLSTM-CRF 模型嵌入层静态向量无法表征多义词问题、编码层未对重要字词进行区分所导致的注意力分散问题,以及缺乏对空间特征捕获的问题,提出了基于 BERT-BiGRU-MHA-CRF 模型和 BERT-IDCNN-CRF 模型集成的中文命名实体识别模型. 利用 BERT 模型综合上下文获得字符的动态表示,解决多义词表征问题;通过多头注意力机制增强对重要字词的关注,解决注意力分散问题;并利用 IDCNN 对输入文本的空间特征进行提取. 最后利用平均集成思想将 BiGRU 网络捕获的时序特征与 IDCNN 捕获的空间特征进行融合,提升模型整体的特征提取能力.

1　本文模型

本文提出的集成命名实体识别模型结构如图 1 所示.

整个模型主要由字符嵌入层(BERT)、编码层(BiGRU-MHA 网络和 IDCNN 网络)、集成层和 CRF 层构成. 中文文本以字符为单位输入 BERT 模型获得包含上下文信息的动态向量表示;接着将向量分别输入 BiGRU-MHA 和 IDCNN 网络获取输入的时序特征和空间特征,其中前者利用多头注意力机制解决编码层注意力分散的问题;然后利用平均集成的方式将两者捕获到的特征进行融合后输入 CRF 层获取全局最优的标注序列. 下文将按照中文命名实体识别流程对各个模块进行分析.

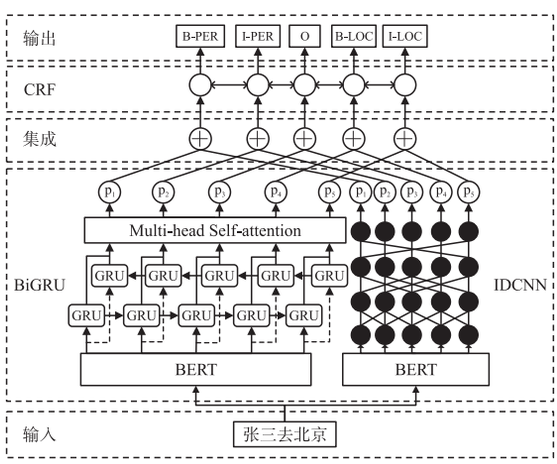

图 1　集成模型结构

Fig. 1　Integrated model structure

1.1　嵌入层

传统的 Word2vec 词嵌入方式仅通过浅层网络训练得到的查找表对输入文本进行简单转换,无法

准确表示中文同形不同义的多义词,而文本能否准确表示直接影响到模型整体的识别效果[8]. 本文使用BERT 预训练模型对输入文本进行动态表示. BERT 模型利用双向 Transformer 编码器对输入字符的上下文信息进行捕获,从而学习到输入中的字符特征和句法特征.

BERT 模型对不同的任务可以有不同的输入形式. 在 NER 任务中,输入一般是具有连续语义的自然文本,每个字符的向量表示都由嵌入向量、分割向量和位置向量组合而成. 其中嵌入向量是关于字符最主要的信息,分割向量用于判断字符属于同一句话,位置向量用于编码每个字符的位置信息,这 3 个向量均在训练过程中通过不断学习得到.

BERT 模型由多层双向 Transformer 编码器构成,它依靠多头自注意力机制实现对输入的上下文建模,Transformer 编码器结构如图 2 所示.

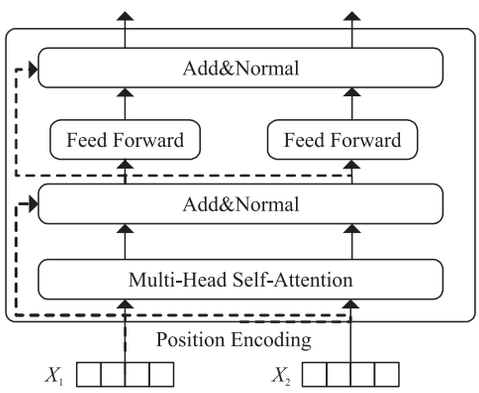

图 2　Transformer 编码单元

Fig. 2　Transformer coding unit

注意力机制是编码器的核心部分,它主要通过计算输入序列中字符之间的关联程度来动态调整权重矩阵,从而获得每个字符在当前语义下的向量表示. 这个新的向量表示不仅包含了该字符本身的含义,还包含了与输入中其他字符的联系,较传统的向量表示方法内容更加丰富合理.

BERT 模型为了提高对于不同位置的专注能力,在 Transformer 编码器中使用了多头自注意力机制(Multi-head Attention,MHA)在多个子表示空间中进行注意力计算. 关于注意力机制的计算过程将在下一小节详细展开论述.

BERT 模型通过注意力机制在对当前字符进行编码时能够关注到序列中其余字符对当前字符的影响,并通过该方式更好地融合上下文信息,提高向量表示的合理性. 正是由于采用了注意力机制才使得BERT 模型能够根据字符所处的上下文环境对其进行动态表示,从而使得本文提出的实体识别模型能够解决中文环境中常见的多义词问题,改善实体识别的效果. 另外根据文献[9]的研究发现,BERT 模型应用于 NER 任务在网络层数为 9 层时效果最好,因此本文的 BERT 模型也采用这一设置.

1.2　编码层

输入文本经过 BERT 层获得每个字符的向量表示,接着传入编码层进行特征提取. 本文提出的集成模型在编码层分别利用 BiGRU-MHA 和 IDCNN 网络对时序特征和空间特征进行提取,接着通过平均集成的方式将提取到的特征进行融合. 在本节中分别对时序和空间特征的提取过程进行详细分析.

1.2.1　BiGRU-MHA 网络

经典模型中的 BiLSTM 网络虽然也能够对输入序列建模,但其单元结构较为复杂,参数较多,训练效率较低. 因此本文利用结构更简单,性能更好的 BiGRU 网络对输入序列的全局特征进行提取.

GRU 网络单元只有更新门和重置门两个门控. 其中的更新门决定了前一时刻的信息有多少传递到当前时刻,而重置门决定了上一时刻的隐藏状态有多少信息需要被遗忘[10]. GRU 单元各个状态计算公式为

$$z_t=\sigma(\boldsymbol{W}_z c_t+\boldsymbol{U}_z h_{t-1}), \tag{1}$$

$$r_t=\sigma(\boldsymbol{W}_r c_t+\boldsymbol{U}_r h_{t-1}), \tag{2}$$

$$\tilde{h}_t=\tanh(\boldsymbol{W}_h c_t+\boldsymbol{U}_h(r_t\odot h_{t-1})), \tag{3}$$

$$h_t=(1-z_t)\odot\tilde{h}_t+z_t\odot h_{t-1}, \tag{4}$$

式(1)-(4)中,z_t,r_t 分别为更新门和重置门,$\tilde{h}_t$ 表示候选隐藏状态,h_t 表示 t 时刻单元的输出,$\odot$ 表示Hadamard乘积,$\boldsymbol{W}_z$,$\boldsymbol{W}_r$,$\boldsymbol{W}_h$ 和 $\boldsymbol{U}_z$,$\boldsymbol{U}_r$,$\boldsymbol{U}_h$ 分别为对应的权重矩阵.

BiGRU 网络的输出也是由前向网络的输出 h_t 和后向网络的输出 h_t 拼接得到,前向网络用于捕获输入的历史信息,后向网络用于捕获后序信息,而拼接后的输出则综合了全局输入信息. BiGRU 层的输出经过全连接层进行维度变换后,会输出一个 $n\times k$ 的状态矩阵,n 和 k 分别为输入序列的长度和标注标签的个数.

BiGRU 网络在提取输入特征时,只是通过门控机制有选择的对当前时刻的上下文进行保留和删除,而没有区分性地看待不同的字符对当前位置的影响. 例如“他的家乡在寒冷的哈尔滨”这句话中,“家乡”对于识别“哈尔滨”这个地名的贡献显然大于“他”“的”这两个字符. 所以本文认为经典模型中在编码层存在注意力分散问题,即对重点字符和普通字符未按重要程度分配不同权重.

为解决这一问题,本文引入了自注意力机制筛选输入文本中的关键信息. 与 Transformer 中的自注意力机制一样,它仅关注输入序列内部字符之间的注意力计算,寻找不同字符之间的联系,筛选出关键的字词信息.

在自然语言处理任务中,常用放缩点积注意力机制,其计算公式为:

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{SoftMax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}, \tag{5}$$

式中,$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别为查询矩阵,键矩阵和值矩阵,d_k 为矩阵 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的维度. 这 3 个矩阵由状态矩阵与随机初始化的权重矩阵相乘所得,在自注意力机制中 $\boldsymbol{Q}=\boldsymbol{K}$. 自注意力机制的计算过程可以分为 3 步,首先计算 $\boldsymbol{Q}$、$\boldsymbol{K}$ 矩阵乘法,并除以 $\sqrt{d_k}$ 防止相乘结果过大. 这一步的目的是为了计算输入中所有字符对于当前字符的重要性;然后经过 SoftMax 归一化操作,将结果归一化为概率分布;最后乘上矩阵 $\boldsymbol{V}$ 表示按照第一步中的重要性更新每个字符的向量表示.

为了提高模型对输入序列中不同位置字符的关注度,本文使用了多头注意力机制. 利用学习到的权重矩阵对 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 进行 h 次不同的线性映射,并行的进行注意力计算,将每个头的结果拼接后再进行一次映射得到的矩阵作为多头注意力层最终的输出. 具体计算公式为

$$head_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}\boldsymbol{W}_i^{V}), \tag{6}$$

$$Multi(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=Concat(head_1,\cdots,head_h)\boldsymbol{W}^{O}. \tag{7}$$

式(6)-(7)中,*Attention* 表示每个注意力头按公式(5)做注意力计算,$\boldsymbol{W}_i^{Q}$,$\boldsymbol{W}_i^{K}$,$\boldsymbol{W}_i^{V}$ 表示第 i 个注意力头中 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 对应的权重矩阵,$\boldsymbol{W}^{O}$ 是最后一次线性映射的权重矩阵.

BiGRU-MHA 网络利用 BiGRU 捕获输入的时序特征并通过注意力机制给每个位置的标签计算得到一组权重值,通过权重值对输出序列加权,筛选输入序列的重要信息. 但其在建模时缺少对局部空间的捕获,而 IDCNN 网络应用于 NLP 任务时能够提取空间特征并消除长期依赖问题,因此本文利用 IDCNN 网络对输入的空间特征进行捕获.

1.2.2　IDCNN 网络

IDCNN 网络是在 Yu F 等人所提 DCNN(dilated convolutional neural network)基础上迭代得到. DCNN 在传统卷积神经网络的基础上加入了膨胀宽度,使得计算时能够跳过膨胀宽度的数据进行卷积,从而在相同卷积核大小的前提下能够获得更宽的上下文信息,增加卷积核的感受范围[11-12]. 卷积核膨胀过程见图 3.

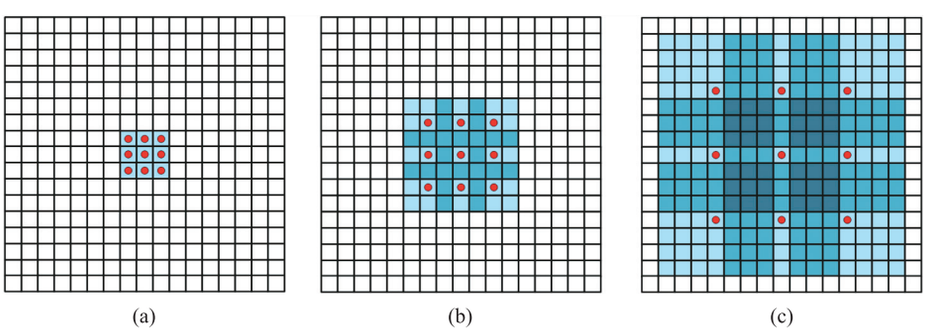

图 3　膨胀卷积示意图

Fig. 3　Dilated convolution diagram

图 3 中 3 个卷积块的膨胀宽度分别为 1,2,4,对应的感受范围分别为 3×3、7×7 和 15×15,在相同卷积核大小前提下,感受范围随着膨胀宽度的增加呈指数增加.

IDCNN 由若干个大小相同的膨胀卷积块堆叠而成,卷积块内部由若干个 DCNN 层组成,且每个卷积

块的输出又作为输出传递给下一卷积块,使得模型能够获得更宽的感受范围,且具有较好的泛化能力. 其具体细节如下.

IDCNN 的第一层利用膨胀宽度为 1 的卷积 $\boldsymbol{D}_1^{(0)}$ 对输入 x_t 进行公式(8)的转换,$\boldsymbol{D}_\delta^{(j)}$ 表示膨胀宽度为 δ 的第 j 层膨胀卷积.

$$i_t=\boldsymbol{D}_1^{(0)}x_t. \tag{8}$$

接着 IDCNN 会使用 L_c 个膨胀卷积层构造出膨胀卷积块,用 *block* $B(\cdot)$ 表示,*block* 块中包含的膨胀卷积如公式(9)-(10),其中 $c_t^{(j)}$ 表示第 j 层的输出.

$$c_t^{(j)}=\mathrm{ReLU}(\boldsymbol{D}_{2^{L_{c-1}}}^{(j-1)}c_t^{(j-1)}), \tag{9}$$

$$c_t^{(L_c+1)}=\mathrm{ReLU}(\boldsymbol{D}_1^{(L_c)}c_t^{(L_c)}). \tag{10}$$

为了在获得更广感受范围的同时避免过拟合,应当避免卷积块过深,而是重复使用 L_b 次相同的卷积块对输入进行处理,具体如公式(11). 初始卷积块的输出为 $b_t^{(1)}=B(i_t)$.

$$b_t^{(k)}=B(b_t^{(k-1)}). \tag{11}$$

最后利用参数 $\boldsymbol{W}_o$ 对最后一层卷积块的输出做公式(12)的映射变换,得到输入序列对于每类标签的得分.

$$h_t{}^{(L_b)}=\boldsymbol{W}_o b_t{}^{(L_b)}. \tag{12}$$

1.3 集成层

当前模型利用 BiGRU-MHA 网络和 IDCNN 网络捕获输入序列的时序和空间特征,但两者所提取到的特征均不够完善,因此在本文模型中加入集成学习思想对所其进行融合,达到同时捕获时序和空间特征的目的.

集成学习是指用某种策略将多种学习器进行结合,从而获得比单个学习器更优的表现. 要达到这一目的,要集成的单一学习器需要具备两个特点:(1)个体学习器必须具备一定的准确性. 例如集成多个二分类学习器时,每个单一的分类器的正确率至少要大于 50%才能保证最终的集成模型分类的效果;(2)要集成的单一模型之间必须具备一定的差异性. 差异性不仅仅是指结果的差异性,更重要的是模型结构和原理上的差异性. 若集成的模型之间毫无差异,则最终的集成模型和单一模型的结果会完全一致,达不到集成的效果.

集成学习中常用的集成策略包括平均法、投票法和学习法[13]. 数值型输出常用简单平均法或加权平均法. 简单平均法将各单一模型输出值的平均作为集成模型的输出;加权平均法是在简单平均法的基础上,根据从训练数据中学到的权重进行加权平均,但由于训练样本存在不充分或噪声问题,其效果并不一定优于简单平均法.

分类任务一般使用标准投票法或加权投票法集成. 标准投票法根据单一模型的分类结果将得票最多的分类类别作为集成模型的输出. 加权投票法通过给每个单一模型分配对应的权重决定对集成模型输出的影响.

学习法是指构建一个新的次级学习器,利用单一模型的输出和初始样本的标记作为一个新的数据集对其进行训练. 而这个次级学习器的输入和输出之间的映射关系就是所要学习的集成策略.

本文使用平均集成法对 BERT-BiGRU-MHA 和 BERT-IDCNN 网络进行集成,并且文献[9]和文献[12]也验证了两者符合集成的必要条件. 在本文模型中,两个网络的输出都通过全连接层映射为 $n\times k$ 的状态矩阵,矩阵中每一行代表当前字符标记为对应 k 个标签的概率. 将两个网络的状态矩阵平均后作为集成层的输出传递到 CRF 层.

1.4 CRF 层

经过编码层对输入序列的建模,模型已经学习到足够的信息对输入序列进行标注,而 CRF 层的主要作用就是对标记序列进行规范,从而得到最优的标记结果.

在 NER 任务中,相邻位置的标注标签往往存在着制约关系. 例如在“O O B-C I-C O”这样一段标注序列中,C 一定代表着同一类实体标签. CRF 层能够在模型训练中根据这样的共现规律排除错误的标注序列,输出全局概率最大的标注结果.

设 $\boldsymbol{P}$ 为集成层输出的融合了时序信息和空间信息的状态矩阵,和 CRF 层中学习到的转移矩阵 $\boldsymbol{M}$ 相

加便得到了输入序列每个字符对应标签的分数 $\boldsymbol{S}(\boldsymbol{X},\boldsymbol{Y})$:

$$\boldsymbol{S}(\boldsymbol{X},\boldsymbol{Y})=\sum_{i=0}^{n}\boldsymbol{M}_{y_i,y_{i+1}}+\sum_{i}^{n}\boldsymbol{P}_{i,y_i}. \tag{13}$$

式(13)中转移矩阵 $\boldsymbol{M}$ 的大小为$(k+2)\times(k+2)$,这是为了增强鲁棒性,在序列的首尾加入了起始和结束标签;$\boldsymbol{P}_{i,y_i}$表示输入序列 X 第 i 个字符标记为 y_i 标签的概率. 则输入为 X 的条件下输出预测序列 Y 的概率为:

$$\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X})=\frac{e^{\boldsymbol{S}(\boldsymbol{X},\boldsymbol{Y})}}{\sum\limits_{\tilde{Y}\in Y_X}e^{\boldsymbol{S}(\boldsymbol{X},\tilde{\boldsymbol{Y}})}}. \tag{14}$$

式(14)中的 $\tilde{Y}$ 表示真实标记序列,Y_X 表示所有可能的标注序列. 利用极大似然法对模型进行训练,使得正确标签的概率最大:

$$\log\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X})=\boldsymbol{S}(\boldsymbol{X},\boldsymbol{Y})-\sum_{\tilde{Y}\in Y_X}\boldsymbol{S}(\boldsymbol{X},\tilde{\boldsymbol{Y}}). \tag{15}$$

最后经过维特比算法得到分数最高的标签序列就是最终 CRF 模型输出的全局最优的标注结果,如公式(16)所示.

$$y^{*}=\underset{\tilde{Y}\in Y_X}{\operatorname{argmax}}\boldsymbol{S}(\boldsymbol{X},\tilde{\boldsymbol{Y}}). \tag{16}$$

2 实验及结果分析

2.1 实验数据及评价指标

为验证所提集成模型的有效性,本文在人民日报 1998 年 1 月新闻标注语料和 MSRA 数据集上进行了实验. 本文将其均按 3∶1∶1 的比例划分为训练集、验证集和测试集,划分之后的具体情况如表 1 所示.

表 1 数据集规模统计

Table 1 Dataset size statistics

数据集	训练集	验证集	测试集
人民日报	801 672	267 314	267 080
MSRA	1 435 571	478 563	478 480

在 NER 任务中,最常用的两类标注体系为 BIO 和 BIOES,本文使用 BIO 标注法. B、I、O 分别表示实体开始字符、实体内容字符和非实体字符.

本文使用精确率 P、召回率 R 和 F_1 值对模型效果进行衡量,其中精确率指识别正确的实体数与识别出实体总数的比值;召回率指正确识别的实体数与实体总数的比值;在某些情况下会出现精确率和召回率冲突的问题,因此利用 F_1 值对 P 和 R 综合考虑. 各指标的计算公式如下:

$$\begin{aligned}P&=\frac{T_p}{T_p+F_p}\times100\%,\\R&=\frac{T_p}{T_p+F_n}\times100\%,\\F_1&=\frac{2PR}{(P+R)}\times100\%,\end{aligned} \tag{17}$$

式中,T_p,F_p 和 F_n 分别为真正例、假正例和假反例的个数.

2.2 实验环境配置

本文具体实验环境如表 2 所示.

表 2 实验环境配置

Table 2 Experimental environment configuration

环境	配置
操作系统	Windows 10
CPU	i7-8700K
GPU	GTX 1660 SUPER
Python	3.6
TensorFlow	1.12.0

2.3 实验参数配置

参考前期研究结果,本文使用网络层为 9 层的 BERT-Base 模型作为集成模型的嵌入层,将 batch_size 设为 128,max_seq_len 为 128,为防止过拟合将 dropout 参数设置为 0.5. IDCNN 网络使用 3×3 大小的卷积核,且卷积层的膨胀宽度分别为 1,1,2.

2.4 实验结果与分析

为研究 IDCNN 卷积核个数和卷积块堆叠层数对实验结果的影响,分别选取 10,20,50,100 个卷积核

和 4,6,8 层卷积块在人民日报数据集上进行实验. 结果分别如图 4,图 5 所示.

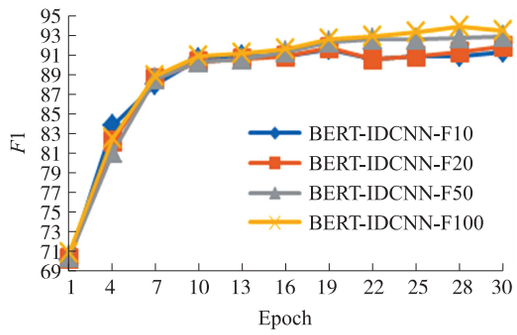

图 4　不同卷积核个数实验结果

Fig. 4　Experimental results of different convolution kernels

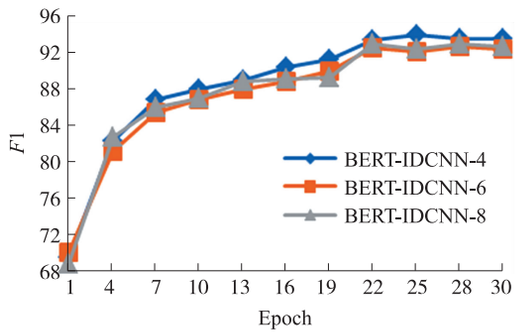

图 5　不同卷积块层数实验结果

Fig. 5　Experimental results of different convolution block layers

图 4 中展示了不同卷积核个数对模型 *F*1 值的影响,其中 *F*10 表示 10 个卷积核. 根据图中结果发现,模型 *F*1 值随卷积核个数的增加有一定的提高,其中卷积核个数为 100 的 BERT-IDCNN-F100 模型效果最好,在第 28 个 epoch 达到了最高的 *F*1 值 93.78%. 图 5 展示了卷积核个数为 100 时不同卷积块层数对模型结果的影响. 随着卷积块层数的增加 *F*1 值逐渐趋于稳定,说明 IDCNN 在 4 个卷积块时已经足够捕获到输入序列的全局信息. 在本文的集成模型中也采用 100 个卷积核,4 个卷积块的设定.

本文集成模型是将 BERT-BiGRU-MHA-CRF 与 BERT-IDCNN-CRF 模型所捕获的特征进行集成. 因此为验证集成模型的有效性,本文也做了单一模型与集成模型在人民日报和 MSRA 数据集上的对比实验,具体结果见表 3 至表 5.

表 3　单一模型在人民日报数据集的结果

Table3　Results of single model in the dataset of People's Daily　%

实体	BERT-IDCNN-CRF			BERT-BiGRU-MHA-CRF		
	P	*R*	*F*1	*P*	*R*	*F*1
人名	95.84	96.46	96.15	98.29	97.51	97.89
地名	94.91	93.72	94.31	96.89	95.33	96.11
机构名	89.62	92.14	90.86	91.25	92.93	92.08
整体	93.46	94.11	93.78	95.58	95.11	95.35

表 4　单一模型在 MSRA 数据集的结果

Table 4　Results of a single model in the data set of MSRA　%

实体	BERT-IDCNN-CRF			BERT-BiGRU-MHA-CRF		
	P	*R*	*F*1	*P*	*R*	*F*1
人名	95.61	96.09	95.85	94.66	96.92	96.79
地名	94.37	93.99	94.18	96.05	95.12	95.58
机构名	89.10	88.95	89.02	87.70	90.82	89.23
整体	92.00	93.20	92.60	94.31	94.61	94.46

表 5　集成模型在两个数据集上的结果

Table 5　The results of the integrated model on the two datasets　%

实体	人民日报			MSRA		
	P	*R*	*F*1	*P*	*R*	*F*1
人名	98.73	97.89	98.31	97.96	98.07	98.03
地名	97.32	95.51	96.41	96.12	95.05	95.58
机构名	93.87	93.62	93.74	90.93	90.42	90.68
整体	96.65	95.53	96.09	95.05	94.97	95.01

观察表 3 和表 4 的实验结果发现,BERT-IDCNN-CRF 模型在人民日报和 MSRA 数据集上整体达到了 93.78%和 92.60%的 *F*1 值,而 BERT-BiLSTM-MHA-CRF 模型则达到了 95.35%和 94.46%的 *F*1 值,说明两者虽然提取到不同的文本特征,但都能够完成中文实体识别,而后者的 *F*1 值略高,说明后者能够更好地对中文文本序列建模,捕获其中包含的关键信息完成输入序列的标注.

但在本文所研究的三类实体中,组织机构类实体的识别效果明显低于其他两类实体,其原因为单个模型的特征提取能力不够,无法较好地对输入序列建模. BERT-IDCNN-CRF 模型在两个数据集上对组织机构实体的 $F1$ 值分别为 9.86%和 89.02%,与其他两类实体有 3%至 7%的差距,而 BERT-BiGRU-MHA-CRF 的结果虽然稍好于前者,但对比来看仍有较大提升空间. 单一结模型的实验结果说明无论是单从时序角度还是空间角度对特征进行提取都无法有效地对组织机构这类嵌套严重的实体进行建模.

表 5 展示了本文集成模型的识别结果. 首先从模型整体的识别效果来看,集成模型在两个数据集上分别达到了 96.09%和 95.01%的 $F1$ 值,同比提高了 0.74%和 0.55%以上. 说明本文集成模型整体上已经达到了比较好的效果.

具体到每一类实体上,集成模型对人名类实体的 $F1$ 值提升了 0.42%以上,地名类实体提升了 0.3%以上,而组织机构名提升了 1.45%以上,提升效果最为明显. 说明集成模型综合时序特征和空间特征后,对不同类型实体的识别效果均有一定程度的提升. 人名类实体和地名类实体结构与机构类实体相比较为简单,原本的单个模型已经能够提取到足够的特征信息,所以在将单个模型集成后整体的提升幅度较小;对于组织机构类实体两个单一模型原本学习到的特征差异较大,在将单一模型集成后能够学习到更为全面的输入序列信息,可以有效地将组织机构这类嵌套严重的实体与人名和地名实体区分开来,提升模型整体的识别效果.

为充分验证本文集成模型的有效性,本文还加入了与目前主流模型在 MSRA 数据集上的对比实验,具体结果见表 6.

表 6　与主流方法的识别效果对比

Table 6　Comparison with the recognition effect of mainstream methods

模型	P/%	R/%	$F1$/%
Lattice-LSTM-CRF[3]	93.57	92.79	93.18
BGRU-CRF[14]	94.65	93.87	94.26
BERT-IDCNN-CRF[12]	94.86	93.97	94.41
BERT-BiGRU-CRF[15]	94.19	94.16	94.18
本文模型	**95.05**	**94.97**	**95.01**

和 Lattice-LSTM-CRF 模型和 BGRU-CRF 模型相比,本文模型引入了裁剪的 BERT 模型作为嵌入层,获得包含上下文的语义向量,并结合注意力机制来增加对输入序列中关键信息的关注度;与 BERT-IDCNN-CRF 模型和 BERT-BiGRU-CRF 模型相比,本文分析了其各自存在的优势并将其各自对于特征抽取的优势进行结合,提高了模型整体的特征捕获能力.

3　结论

本文针对命名实体识别经典 BiLSTM-CRF 模型嵌入层的静态向量无法表征多义词问题、编码层未对重要字词进行区分所导致的注意力分散问题,以及缺乏对空间特征捕获的问题,提出了基于 BERT-BiGRU-MHA-CRF 模型和 BERT-IDCNN-CRF 模型的集成模型. 并在人民日报和 MSRA 数据集上进行实验验证,分别取得了 96.09%和 95.01%的 $F1$ 值,相较于单个模型至少提升了 0.74%和 0.55%的整体 $F1$ 值,并且由于集成模型能够学习到更为全面的输入序列特征,因此对于组织机构这类嵌套严重的中文实体识别效果提升明显,$F1$ 值提升达到了 1.45%以上. 本文仅对通用领域的部分类型的实体进行了研究,下一步工作可以扩展到细粒度或其他垂直领域实体的识别中去,以验证本文所提集成模型的有效性和在不同领域实体识别上的普适性.

[参考文献]

[1] HAMMERTON J. Named entity recognition with long short-term memory[C]//Proceedings of the Seventh Conference on Natural Language Learning at Annual Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies 2003. 2003:172-175.

[2] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[J]. arXiv preprint arXiv:1603.01360,2016.

[3] ZHANG Y, YANG J. Chinese NER using lattice LSTM[J]. arXiv preprint arXiv:1805.02023,2018.

[4] GUI T, MA R, ZHANG Q, et al. CNN-Based Chinese NER with Lexicon Rethinking[C]//Twenty-Eighth International Joint Conference on Artificial Intelligence. Macao:Springer,2019:4982-4988.

[5] WANG C Q, CHEN W, XU B. Named entity recognition with gated convolutional neural networks[M]//Chinese Computational

Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer,Cham,2017:110-121.
[6] YU F,KOLTUN V. Multi-scale context aggregation by dilated convolutions[J]. arXiv preprint arXiv:1511.07122,2015.
[7] YU B H,WEI J X. IDCNN-CRF-based domain named entity recognition method[C]//2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology(IEEE 2nd International Conference on Civil Aviation Safety and Information Technology:542-546.
[8] 杨晓辉,毕雪华,张琳琳,等. 基于多任务的中文电子病历中命名实体识别研究[J]. 东北师大学报(自然科学版),2020,52(1):81-87. DOI:10.16163/j.cnki.22-1123/n.2020.01.016.
[9] 孙弋,梁兵涛. 基于 BERT 和多头注意力的中文命名实体识别方法[J/OL]. 重庆邮电大学学报(自然科学版):1-10[2022-02-13]. http://kns.cnki.net/kcms/detail/50.1181.N.20211209.2010.004.html.
[10] 张柯文,李翔,严云洋,等. 基于多特征双向门控神经网络的领域专家实体抽取方法[J]. 南京师大学报(自然科学版),2021,44(1):128-135.
[11] 孔祥鹏,吾守尔·斯拉木,杨启萌,等. 基于迁移学习的维吾尔语命名实体识别[J]. 东北师大学报(自然科学版),2020,52(2):58-65. DOI:10.16163/j.cnki.22-1123/n.2020.02.010.
[12] 李妮,关焕梅,杨飘,等. 基于 BERT-IDCNN-CRF 的中文命名实体识别方法[J]. 山东大学学报(理学版),2020,55(1):102-109.
[13] 周志华. 机器学习[M]. 北京:清华大学出版社,2016.
[14] 石春丹,秦岭. 基于 BGRU-CRF 的中文命名实体识别方法[J]. 计算机科学,2019,46(9):237-242.
[15] 杨飘,董文永. 基于 BERT 嵌入的中文命名实体识别方法[J]. 计算机工程,2020,46(4):40-45,52. DOI:10.19678/j.issn.1000-3428.0054272.

[责任编辑:顾晓天]