doi:10.11887/j.cn.202101012 http://journal.nudt.edu.cn

中文分词与命名实体识别的联合学习*

黄晓辉[1,2],乔立升[1],余文涛[2],李 京[1],薛 寒[2]
(1. 中国科学技术大学 计算机科学与技术学院,安徽 合肥 230026;
2. 战略支援部队信息工程大学洛阳校区,河南 洛阳 471003)

摘 要:将卷积结构引入循环神经网络,从而构建卷积循环神经网络。以此为基础,研究构建了面向中文分词与实体识别联合学习的序列标注模型。该模型依托卷积循环神经网络构建特征编码层,实现中文字序列局部空间特征和长距离时序依赖特征的联合提取;依托改进的循环神经网络构建标签解码层,实现标签序列长距离时序依赖的有效建模;依托统一的分词与实体识别序列标注模式实现分词信息与实体信息的联合学习,避免传统流水线法的误差传播问题。在人民日报语料和微软标注语料上的实验结果显示,该框架较传统统计模型和神经网络模型有显著的性能提升,尤其是在识别字数较多的命名实体时,其效果明显优于其他方法。

关键词:卷积循环神经网络;局部空间特征;时序依赖特征;分词与实体识别
中图分类号:TP183 **文献标志码**:A **文章编号**:1001-2486(2021)01-086-09

Joint learning of Chinese word segmentation and named entity recognition

HUANG Xiaohui[1,2], *QIAO Lisheng*[1], *YU Wentao*[2], *LI Jing*[1], *XUE Han*[2]
(1. College of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China;
2. Luoyang Campus of the Information Engineering University of the Strategic Support Force, Luoyang 471003, China)

Abstract: The convolutional structure was introduced into the recurrent neural network to construct a convolutional recurrent neural network. Based on this network, a sequence annotation model for joint learning of Chinese word segmentation and entity recognition was constructed. The model relies on the convolutional recurrent neural network to construct feature-encoding layer, which realizes the joint extraction of local spatial features and long-distance time-dependent features of Chinese character sequences; the improved recurrent neural network was relies on the constructing of tag-decoding layer, which realizes the effective modeling of timing-dependent features in the tag sequences; the unified word segmentation and entity recognition annotation mode relies on the achieving of joint learning of word segmentation information and entity information, which avoids the error propagation problem of traditional pipeline methods. Experimental results on the People's Daily corpus and Microsoft's annotated corpus show that the framework has significant performance improvement over traditional statistical models and neural network models, especially when identifying entities with multiple characters, and its effect is significantly better than other methods.

Keywords: convolutional recurrent neural network; local spatial features; time-dependent features; word segmentation and entity recognition

命名实体识别也称专名识别,旨在识别出文本中表示命名实体的成分,是篇章理解、信息检索、知识图谱、机器翻译等自然语言处理高层应用的基础[1]。解决命名实体识别的主流方法是将其作为序列标注问题,即为输入文本序列中的每个字(词)预测一个标签,该标签包含了实体的边界信息和类别信息。目前,主要有基于统计的序列标注模型和基于神经网络的序列标注模型两大类。基于统计的序列标注模型从概率的角度来建模输入序列与标签序列之间的关系,如隐马尔可夫模型(Hidden Markov Model, HMM)、条件随机场模型(Conditional Random Field, CRF)等[2]。例如,文献[3]设计了基于层叠HMM的中文命名实体识别模型,文献[4]则基于CRF构建面向生物医学文本的命名实体识别模型,有效提升了领域实体的识别性能。

近年来,得益于强大的特征提取能力和非线性拟合能力,基于神经网络的序列标注模型逐渐崭露头角。与统计模型相比,神经网络模型对特征选取的依赖程度大大降低[5],在命名实体识别

* 收稿日期:2019-08-27
基金项目:国家重点研发计划资助项目(2016YFB0201402)
作者简介:黄晓辉(1986—),男,河南洛阳人,讲师,博士研究生,E-mail:huangxia@mail.ustc.edu.cn

领域获得了显著的性能提升,如文献[6]采用前馈神经网络,结合字词特征在中文新闻语料上取得了非常好的效果。文献[7-8]则分别基于循环神经网络(Recurrent Neural Network, RNN)的两种变体——双向长短时记忆(Bi-directional Long Short-Term Memory, Bi-LSTM)网络[9]和网格长短时记忆网络来设计序列标注模型,并成功应用于中文命名实体识别。此外,业界还提出通过组合模型来实现命名实体识别,如文献[10]将 Bi-LSTM 与 CRF 相结合,以 LSTM 来提取序列特征,以 CRF 建模标签的时序依赖,在英文命名实体识别上取得了显著的性能提升。文献[11]则利用卷积神经网络(Convolutional Neural Network, CNN)来强化模型的特征提取能力,构建了基于 CNN、LSTM 以及 CRF 的序列标注模型,并将其应用于生物医学领域的命名实体识别。文献[12]则使用 LSTM-CNNs-CRF 的组合架构进行序列标注,同样在命名实体识别上获得了显著进步。总体上,组合结构模型旨在利用不同组件实现不同特征的提取。现有组合模型中各组件通常在结构上相互独立,在数据处理过程上又顺序依赖,模型整体复杂度增加,不仅给模型训练带来诸多问题,也给局部空间特征和时序依赖特征的联合提取带来了较大的不确定性。

同时,由于中文的最小语义单位是汉字,因此可基于字标注方法实现命名实体识别。但在实际应用中,由于字的语义信息太过简略,因此通常会采用先分词、再基于词序列标注的流水线处理模式。这种模式下,分词误差不可避免地会影响后续实体识别的效果。

基于以上现状,本文从三方面开展研究工作:一是针对中文字序列局部空间特征和时序依赖特征的联合提取问题,研究构建融合卷积结构和循环结构的神经网络特征提取器,以实现序列数据局部空间特征和长距离时序依赖特征的联合建模;二是针对标签序列上下文特征的提取问题,研究构建基于改进 RNN 的标签解码网络,使模型在预测标签时能够充分利用标签序列的上下文关联特征;三是针对传统流水线模式带来的误差传播问题,研究设计融合词边界信息和实体信息的字序列标注模式,将中文分词与实体识别纳入统一的联合学习框架,在有效利用分词信息辅助命名实体识别的同时,避免分词误差的传播问题。与现有主流命名实体识别方法相比,本文提出的模型具有三个显著特点:一是基于卷积循环神经网络构建特征编码器,充分发挥两者的优势来实现序列数据局部空间特征和时序依赖特征的有效提取;二是基于改进 RNN 的标签解码网络,充分利用 RNN 对时序特征的建模能力,将标签序列的上下文信息纳入学习过程;三是统一的分词与实体识别序列标注模式较传统标注模式纳入了更多的信息,借助于模型强大的学习能力,实现中文分词与实体识别的联合学习。最终实验结果证明,该联合学习模型对中文命名实体有更好的识别能力,尤其是在识别字数较多的命名实体时,其效果要明显优于其他方法,这也是本文最突出的贡献点。

1　中文分词与实体识别联合学习框架

本文提出的基于融合结构神经网络的中文分词与实体识别联合学习模型总体架构如图 1 所示。

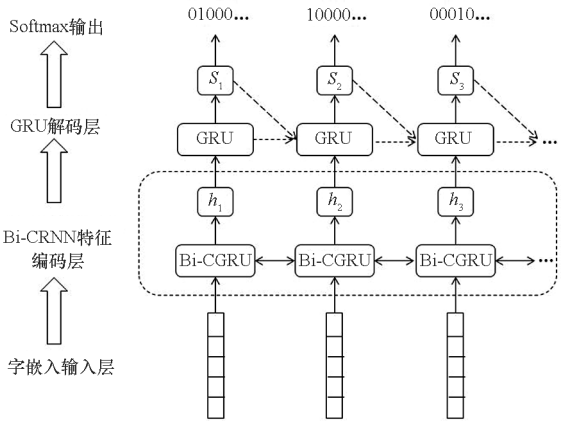

图 1　基于融合结构神经网络的中文分词与实体识别联合学习模型

Fig. 1　Joint learning model of Chinese word segmentation and entity recognition based on fused neural network

该模型主要由四部分构成:一是由中文字向量构成的输入层,以实数向量编码的方式实现汉字之间语义关系的初步建模,称为字嵌入层。二是融合卷积和循环结构的特征编码网络,称为 Bi-CRNN 编码层。该层借助于卷积循环神经网络强大的特征提取能力,实现输入序列局部空间特征和长距离时序依赖特征的联合编码。三是基于改进的循环神经网络构建的标签解码层,称为 GRU 解码层。该层以特征编码层输出向量为前馈输入,经隐层变换形成隐状态向量,再将前一时刻输出的标签概率向量与特征编码向量整合,从而综合利用特征编码信息和标签上下文信息解码出标签序列。四是 Softmax 分类层,所表示的标签结构将字在词中的位置信息、字在

实体中的位置信息以及实体的类型信息进行统一编码,作为分词与实体识别联合学习的依据。最终,结合特定的目标函数和训练算法对模型进行训练,即可实现中文分词与实体识别的联合学习。

2 基于卷积循环神经网络的特征编码层

为实现序列数据局部空间特征和时序依赖特征的联合建模,本文将 CNN 的局部连接、权值共享结构引入 RNN 的前馈连接中,同时对神经元的工作模式进行相应改进,在保留时序特征提取能力的同时,改善其对局部空间特征的提取效果。经过改进的神经网络称之为卷积循环神经网络,网络中每个神经元都是一个具有时序自连接的特征 filter,通过在输入序列上做二维卷积来提取局部空间特征,再做一维的时序迭代来提取上下文时序特征。卷积循环层中单个神经元的数据处理过程如图 2 所示。

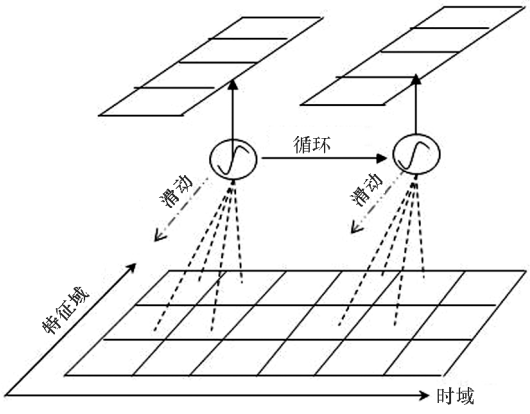

图 2 单个卷积循环神经元的特征提取过程

Fig. 2 Feature extraction process of a single convolutional recurrent neuron

图 2 中,单个特征 filter 以一定大小的感知区域(称为 patch,图中为 2 × 2 区域)先沿特征维作一维卷积操作(图中“ -·-·”),从而生成某一时刻的局部特征 map(实际上是一个向量),之后再沿时间维滑动,在下一时刻先对序列数据做同样的卷积操作,再利用时序连接将前一时刻生成的局部 map 引入新的局部特征 map 中,从而实现对局部空间特征和时序依赖特征的联合提取。

由于特征编码层存在多个卷积循环神经元,因此每个神经元在某一时刻都会产生瞬时的局部特征 map,多个神经元就会产生多个特征 map,进而在时序迭代时需要对前一时刻的多个特征 map 进行处理,因此本文对神经元的时序迭代过程进行了改进,以两个卷积循环神经元的数据处理流程为例,其过程如图 3 所示。

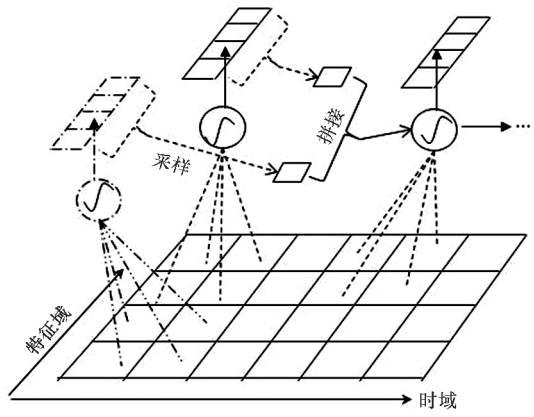

图 3 两个卷积循环神经元的特征提取过程

Fig. 3 Feature extraction process of two convolutional recurrent neurons

图 3 中,在某一时刻,两个神经元(图中第一时刻“----”表示的 kernel 和 “-··”表示的 kernel)生成了两个局部瞬时特征图。在时序迭代之前,对每个特征 map 进行最大值采样(图中短划线型虚线大括号及箭头),只保留 map 中的最大值;之后将同一时刻生成的所有特征图的采样值进行拼接(图中实线大括号),形成当前时刻的隐状态特征向量;最后该特征向量经加权后进入下一时刻的卷积循环神经元。在这种处理方式下,前一时刻生成的隐状态向量维数仍然是神经元的个数,因此循环连接的数量仍为 $N \times N$(N 为卷积循环神经元的个数),可直接依据 RNN 的自连接运算方式进行时序迭代。卷积循环层的计算过程如下:

$$\begin{cases} m_{t,n} = c_{\text{onv1}}(x_t, C_n) \\ y_{t+1,n} = R_{\text{nncell}}(m_{t+1,n}, [p_{\text{ool}}(m_{t,1}), \cdots, \\ \qquad p_{\text{ool}}(m_{t,N})]) \\ \boldsymbol{y}_{t+1} = [y_{t+1,1}, \cdots, y_{t+1,N}] \end{cases} \tag{1}$$

式中,$\boldsymbol{x}_t$ 表示输入序列在 t 时刻的向量,$m_{t,n}$ 表示在 t 时刻由第 n 个神经元 C_n 在向量 $\boldsymbol{x}_t$ 的上下文环境中经过 1 维卷积生成的瞬时特征 map,$y_{t+1,n}$ 表示在 $t+1$ 时刻第 n 个循环卷积神经元生成的特征编码值,其以 $t+1$ 时刻的瞬时卷积特征 map 作为前馈输入,以 t 时刻所有神经元生成的特征 map 经过 Pooling 并拼接(公式中中括号即表示拼接)后形成的向量作为循环输入,最终由 RNN cell 生成一个特征编码值。最后,$t+1$ 时刻所有神经元的特征编码值经过拼接形成 $t+1$ 时刻的特征编码向量 $\boldsymbol{y}_{t+1}$。

借鉴循环神经网络中的门控机制[13]和双向时序迭代机制[14],本文基于卷积循环神经网络构建了特征编码层,其对数据的处理流程如图 4 所示。

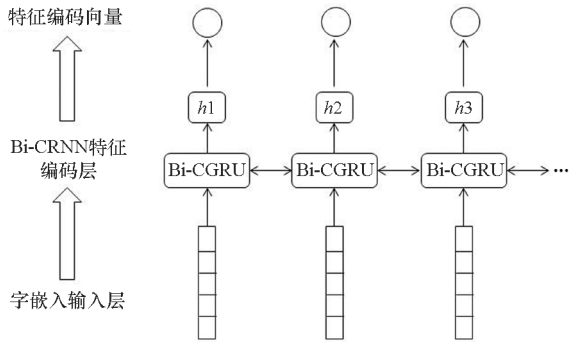

图 4　特征编码层的数据处理过程示意

Fig. 4　Data processing process of the feature encoding layer

图 4 中,编码层的双向迭代机制的 CGRU 网络(Bi-directional Convolutional Gated Recurrent Unit, Bi-CGRU),即设置两个卷积循环神经层,分别从输入序列的两个方向进行局部空间特征和时序依赖特征的提取。CGRU 则表示基于 GRU cell 门控机制的卷积循环神经网络,即采用 GRU 门控机制来对卷积操作生成的特征 map 进行处理,其过程可如图 5 所示。

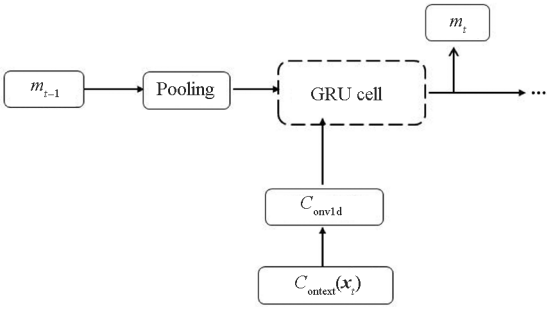

图 5　CGRU Cell 的数据处理过程示意

Fig. 5　Data processing process of the CGRU Cell

最终,所有神经元在同一时刻生成的多个 map 在纵向形成特征编码向量,在横向经 Pooling 以及拼接后,形成时序迭代的状态向量,进入下一时刻的神经元中。由于就某一时刻输入字向量而言,编码网络中两个方向的卷积循环层会生成两个编码向量序列,因此将同一时刻两个方向的特征编码向量进行拼接,形成一个对该时刻输入向量的总体特征描述。

3　基于改进 RNN 的标签解码层

为了建模标签序列的长距离时序依赖关系,本文在特征编码层之上通过改进 RNN 构建了标签解码层,其运行过程如图 6 所示。

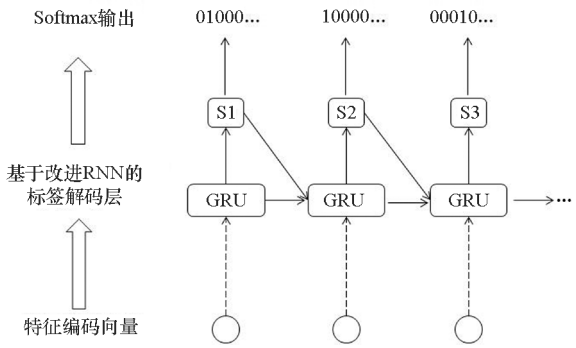

图 6　标签解码层的运行过程示意

Fig. 6　Operation process of the tag decoding layer

该解码层是对传统 RNN 的改进,其门控机制仍然采用 GRU cell,在保留隐含层时序自连接的同时,还将前一时刻 Softmax 分类网络的输出值引入当前的隐含层状态中,不仅利用隐含层循环连接来建模特征编码的时序依赖,还利用前一时刻输出标签的信息来辅助当前时刻标签的预测。基于改进 RNN 的标签解码层内部结构及数据处理流程如图 7 所示。

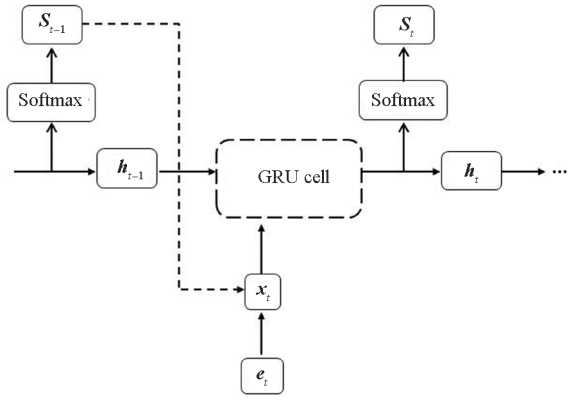

图 7　GRU 解码网络内部结构及数据处理流程

Fig. 7　Internal structure and data processing flow of the GRU decoding network

从图 7 可以看出,该标签解码层的门控单元与传统的 GRU 网络是一致的,只是在数据进入 GRU cell 时加入了一个额外的处理步骤,即将前一时刻 Softmax 分类层的输出值与特征编码层的编码向量进行了拼接,之后一起作为 GRU 的前馈输入,来实现当前时刻的标签预测。其计算过程如式(2)所示:

$$\begin{cases} \boldsymbol{x}_t = [\boldsymbol{S}_{t-1}, \boldsymbol{e}_t] \\ \boldsymbol{h}_t = \boldsymbol{G}_{\text{RUcell}}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}) \\ \boldsymbol{S}_t = \boldsymbol{S}_{\text{oftmax}}(\boldsymbol{h}_t) \end{cases} \tag{2}$$

式中,$\boldsymbol{e}_t$ 表示特征编码层在 t 时刻输入的编码向

量，$\boldsymbol{h}_t$ 表示解码网络隐含层状态向量，$\boldsymbol{S}_t$ 表示解码网络在 t 时刻输出的标签类别概率向量。可以看出，在 t 时刻，编码层的特征向量与前一时刻的标签类别概率向量进行拼接后，进入 GRU cell 进行运算处理后形成隐状态向量，其中一路进入分类层输出 t 时刻的标签概率，另一路进入下一时刻的 GRU cell 进行时序迭代。

4 统一的分词与实体识别序列标注模式

为实现联合学习，本文设计了融合词边界信息、实体边界信息和实体类别信息的标签结构。该标签结构共由三个子部分构成：第一部分为分词标记，表示字在词中的位置；第二部分为实体的类别标记，表示该字所属实体的类别（本文采用人名、地名和组织机构名三个类别作为示例）；第三部分为实体边界标记，表示该字在实体中的位置。标签结构以及所代表的含义如表 1 所示。

表 1 分词与实体识别联合学习标签结构

Tab. 1 Tag structure for the joint learning of word segmentation and entity recognition

字在词中的位置	字所属实体类别	字在实体中的位置
词首(B)	人名(PER)	实体首字(B)
词中(M)	地名(LOC)	实体中字(I)
词尾(E)	机构名(ORG)	实体尾字(L)
单字词(S)	其他(O)	单字实体(S)

本文设计的标签模式将字在词中的位置、字在实体中的位置以及字所属实体的类型进行了整合，形成统一的中文分词与实体识别序列标注模式。理论上讲，所有标签类别应有 $4 \times 4 \times 4 = 64$ 种，但实际上并没有这么多。从语法规则来讲，有些组合是不存在的（例如，位于词中间的字不可能是一个实体的开始或结尾）。最终，结合语法规则以及对语料库中标签的筛选，在去除不合理、不存在的标签组合后，共得到如表 2 所示的标签集合。

由于各标签之间互斥存在，因此本文采用 one-hot 向量形式对标签进行编码，共设置 31 维标签编码向量（表示 31 个标签类别，每一维代表一个类别），标签编码向量中只有一个维度的值为 1，其他维度的值则为 0。

表 2 分词与实体识别联合学习标签集

Tab. 2 Label set for the joint learning of word segmentation and entity recognition

标签类别	标签集合
人名实体标签	B-PER-B, B-PER-I, M-PER-I, E-PER-I, E-PER-L, S-PER-B, S-PER-I, S-PER-L, S-PER-S
地名实体标签	B-LOC-B, B-LOC-I, M-LOC-I, E-LOC-I, E-LOC-L, S-B-LOC, S-LOC-I, S-LOC-L, S-LOC-S
机构实体标签	B-ORG-B, B-ORG-I, M-ORG-I, E-ORG-I, E-ORG-L, S-ORG-B, S-ORG-I, S-ORG-L, S-ORG-S
非实体标签	B-O, M-O, E-O, S-O

5 实验及结果分析

5.1 实验任务及数据准备

为验证所设计框架的有效性，本文分别以人民日报标注语料（PFR）和微软标注语料（MicroSoft Research Asia, MSRA）为基础数据进行试验验证，两个语料库的具体情况如下。

98 版人民日报标注语料：来自 1998 年 1—6 月的人民日报新闻语料，其中新闻文本进行了分词和词性标注。语料中每个句子由换行符隔开，句子中词与词之间由空格隔开，每个词后面会跟一个词性标记。命名实体信息包含于词性信息中，其中标记 n_r，n_t，n_z 分别代表人名、地名和组织机构名。据统计，语料库中命名实体的信息如表 3所示。

表 3 PFR 标注语料的统计信息

Tab. 3 Statistics of the PFR

实体类别	单字实体	多字独词实体	多词实体	总数
人名实体	4 210	28 250	88 161	120 621
地名实体	0	133 014	7 256	140 270
机构实体	0	20 079	54 392	74 471
总体	4 210	181 343	149 809	335 362

微软亚洲研究院命名实体标注语料库：由微软亚洲研究院提供，也是 Sighan2006 backoff 3 使用的中文语料库。该语料库根据用途划分为分词版本和命名实体标注版本。两个版本的语

料内容是一致的,只是标记不同。该语料库的实体统计信息如表 4 所示。

表 4　MSRA 标注语料的统计信息

Tab. 4　Statistics of the MSRA

	实体类别	单字实体	多字实体	总数
训练集	人名实体	962	16 653	17 615
	地名实体	5 997	30 520	36 517
	机构实体	91	20 480	20 571
	总体	7 050	67 653	74 703
测试集	人名实体	61	1 912	1 973
	地名实体	445	2 432	2 877
	机构实体	7	1 324	1 331
	总体	513	5 668	6 181

在进行实验之前,对这两个语料库中的句子和标记进行了预处理。对于 PFR 语料库,将分词和实体标记转换为基于字的联合标记,随机选取 80% 的句子作为训练集和 5% 的句子作为验证集以及 15% 的数据作为测试集。对于 MSRA 语料库,依据分词版本和实体标注版本中语料的对齐关系,将分词标记和实体标记转换为联合标记,同时由于该语料库已经预先划分了训练集和测试集,因此本文从训练集中随机划分出 10% 的语料作为验证集。

5.2　模型构建与训练

依据图 2 所示的联合学习框架,采用 python 3.5 + Tensorflow 1.4 来构建序列标注模型。其中,输入层设置 100 个节点,表示 100 维的特征向量,本文采用 word2vec 框架在 PFR 和 MSRA 语料上一起训练得到 100 维的中文字向量作为输入;特征编码层和标签解码层的参数在实验过程中根据模型的训练和测试效果来确定;Softmax 分类层则设置 31 个节点,代表 31 维的标签向量输出。模型训练采用的目标函数为样本集的 Log 似然,以最大化该似然函数作为训练目标,其定义如式(3)所示:

$$LL(\theta) = \sum_{j=1}^{|D|}\sum_{t=1}^{L_j}\left[\log_2(p(c_t^j = y_t^j))\right] \quad (3)$$

式中,$|D|$表示整个训练集中的文本句子序列及其标注序列,L_j 表示编号为 j 的句子的长度,c_t^j 是标注模型为句子 j 中第 t 个字预测的标签类别,y_t^j 为句子 j 第 t 个汉字的真实标签,$p(c_t^j = y_t^j)$ 表示模型预测得到真实标签的概率。模型最终的训练目标是得到一组参数 θ,该参数能够使该似然函数达到最大值,即:

$$\theta^* = \arg\max_{\theta} LL(\theta) \quad (4)$$

模型训练时,本文首先根据句子的长度对所有句子进行升序排序,之后再根据句子长度所在的区间进行了统一的长度 Padding,即设置[0, 15]、[16, 25]、[26, 35]三个区间,长度落在相应区间内的句子统一 Padding 到所在区间的右边界值(补 0)。在训练过程中,随机选取长度在某个区间的所有句子作为一个大的样本集合,之后在该集合内再采用基于 Minibatch 的随机梯度下降算法及其变体来训练模型。其中,Minibatch 的值设置为 32;模型优化算法选择 Adam 算法[15],采用 Tensorflow 默认的参数配置($learning\_rate = 0.001$, $beta1 = 0.9$, $beta2 = 0.999$, $epsilon = 1E-08$);模型初始权重由均值为 0、方差为 1 的标准正态分布在[-1,1]区间内随机生成;每当训练集上完成一轮迭代,就在验证集上进行一次验证。当模型在验证集上的性能趋于稳定或是开始持续下降时,模型停止训练。

训练完成后,将在测试集上对模型进行性能评估。由于最终输出是概率向量的序列,因此在进行实体切分及类型判断时,对该向量序列分两步进行处理:首先去除不包含实体信息仅有分词信息的标签;之后依据标签采用最近配对原则切分出实体的边界,并判断出其类型。最终,根据识别出的命名实体情况,采用 Precision、Recall、F1 - 值三个量作为指标,用于评价模型最终表现,其定义如式(5)所示。

$$\begin{cases} \text{准确率(Precision)} = \dfrac{\text{正确识别出的实体个数}}{\text{识别出的实体总数}} \times 100\% \\ \text{召回率(Recall)} = \dfrac{\text{识别出的实体个数}}{\text{实体的总数}} \times 100\% \\ \text{F1 - 值} = \dfrac{2 \times \text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}} \end{cases} \quad (5)$$

5.3　实验结果分析

在模型训练过程中,本文重点对编码网络和解码网络的超参数设置进行了多次试验验证,以探索不同结构对标注性能的影响,如神经元节点个数、卷积 patch 的大小、卷积 stride 的大小等,以标注效果最好的结构作为最终的模型。最终获得的模型结构及在数据集上的标注结果如表 5 和表 6所示。

表5　识别性能最佳的模型参数

Tab.5　Parameters of the best performance

网络层	结构类型	具体参数
输入层	字向量嵌入层	100维(word2vec)
编码层	双向卷积循环层	双向各256个卷积循环神经元,每个神经元的卷积patch为5×5
解码层	改进的RNN网络	单向128个改进的GRU cell
分类层	Softmax层	31维

表6　模型在测试集上的实体识别结果

Tab.6　Entity recognition results on the test set

数据集	识别对象	准确率/%	召回率/%	F1-值/%
PFR	人名实体	94.42	95.21	94.81
	地名实体	95.15	93.48	94.31
	机构实体	94.22	92.22	93.21
	总体	94.68	93.82	94.25
MSRA	人名实体	93.12	93.67	93.39
	地名实体	94.32	92.28	93.28
	机构实体	92.92	91.22	92.06
	总体	93.62	92.49	93.05

从表6可以看出,本文设计的联合学习模型在两个数据集上都取得了较好的识别效果,各个指标都在90%以上。尤其是对于组织机构名的识别,其结果与人名的识别结果已基本相当,这在传统基于序列标注模型的命名实体识别任务中是很少见的。因为传统命名实体识别模型通常对人名、地名有较高的识别率,而对组织机构名的识别效果要差很多,其原因就在于人名一般较短、较简单,而地名一般具有较明显的指示特征,因而通常有较高的识别率。而组织机构名不仅字数较多,并且可以包含人名和地名,其内容形式极其丰富,有着复杂的局部空间特征和长距离时序依赖特征。因此,一般模型对这种数据特征模式难以产生最佳的提取效果,而本文提出的网络模型正是针对这一特征模式的有效提取而设计,因此才取得了更好的识别效果。

此外,为了验证所设计的联合学习框架较其他模型在中文命名实体识别方面的优越性,本文还与业界公认性能较好的基于CRF、LSTM、CNN及组合结构的序列标注模型进行了对比实验,同时也与目前公开发表的、在PFR语料库上效果较好的研究工作进行了对比,以三类命名实体(人名、地名、组织机构名)的准确率、召回率和F1-值作为参考指标,最终详细的对比结果如表7所示。

表7　与其他模型的识别性能对比(PFR)

Tab.7　Comparison with other models' recognition performance(PFR)

模型架构	识别对象	准确率/%	召回率/%	F1-值/%
CRF/词标注	人名实体	93.10	90.75	91.91
	地名实体	95.15	91.38	93.23
	机构实体	83.16	82.21	82.68
Bi-LSTM+Softmax/词标注	人名实体	91.78	91.62	91.70
	地名实体	94.24	93.12	93.68
	机构实体	82.23	83.16	82.69
Bi-LSTM+CRF/字标注	人名实体	93.98	94.47	94.22
	地名实体	95.42	92.36	93.87
	机构实体	84.32	85.62	84.97
CNN+LSTM+CRF/字标注	人名实体	94.66	95.32	94.99
	地名实体	94.42	93.76	94.09
	机构实体	86.42	84.41	85.40
张海楠,伍大勇等[6]	人名实体	92.58	91.97	92.27
	地名实体	93.20	96.03	94.59
	机构实体	87.22	89.96	88.57
CRNN+RNN/字标注联合学习	人名实体	94.42	95.21	94.81
	地名实体	95.15	93.48	94.31
	机构实体	94.22	92.22	93.21

从表7的结果来看,本文设计的基于"CRNN+RNN/字标注联合学习"模型在三个指标上获得了最佳的识别结果(组织机构名的准确率、召回率和F1-值都排名第一),在五个指标上获得了第二的结果(与排名第一的差距都在0.5%以内),因此总体上较其他模型有显著的性能提升。同时与文献[6]发表的识别结果相比,本文提出的模型尽管在地名识别方面稍有落后,但在人名和组织机构名的识别方面有明显的性能提升。并且,文献[6]提出的方法使用了大量精准的字词特征,而本文提出的方法没有涉及任何的外部特征。另外,在实验过程中,本文还对各个模型的参数数量、训练过程迭代次数以及标注效率进行了统计和对比分析,结果显示,本文所提模型具有最少的训练参数,且在训练迭代次数和序列解码时间上都要明显少于其他神经网络模型,因此较其他模型具有更高的训练和推理效率。

另外,本文基于改进RNN构建的标签解码层

是与传统模型的重要不同之处。为验证该标签解码层对标注效果的影响,本文以 CRNN 编码层为基础,针对 Softmax 层、CRF 解码层和基于改进 RNN 的解码层进行了对比实验,其结果如表 8 所示。

表 8　不同解码层的性能对比(PFR)

Tab. 8　Performance comparison of different decoding layers(PFR)

模型结构	准确率/%	召回率/%	F1 - 值/%
CRNN + Softmax	91.84	91.62	91.73
CRNN + CRF	94.76	92.98	93.86
CRNN + RNN	94.68	93.82	94.25

从表 8 可以看出,在同样采用 CRNN 作为特征编码层的情况下,配置 CRF 解码层的模型较单纯的 Softmax 分类层有更高的准确率和召回率以及 F1 - 值,这说明 CRF 层较 Softmax 层对标签上下文信息的建模能力要强得多,同时也说明,标签的上下文关联信息对于命名实体识别的结果有着直接的影响;同样的情形,基于改进 RNN 的解码层较基于 CRF 构建的解码层在具有相当准确率的情况下,其召回率又有了明显提升,这说明有更多的命名实体被识别出来,并且是被准确地识别出来,证明了基于改进 RNN 的解码层较 CRF 层对标签的上下文关联特征有更好的建模能力,因而提升了命名实体的识别效果,验证了这一设计的有效性。

此外,为了验证模型在识别长实体时的优越性,本文专门针对字数超过 6 的命名实体(主要是地名和组织机构名)识别结果进行了统计对比,其结果如表 9 所示。

表 9　对长实体的识别性能对比(PFR)

Tab. 9　Comparison of recognition performance of long entities(PFR)

模型架构	准确率/%	召回率/%	F1 - 值/%
CRF/词标注	83.16	82.21	82.68
Bi-LSTM + Softmax/词标注	82.57	84.64	83.59
Bi-LSTM + CRF/字标注	86.19	84.55	85.36
CNN + Bi-LSTM + CRF/字标注	88.30	87.68	87.98
CRNN + RNN/词标注	91.92	90.12	91.01
CRNN + RNN/字标注联合学习	93.96	92.08	93.01

从表 9 可以看出,对于具有多个字的长命名实体的识别,本文设计的联合学习模型较其他模型有明显的优势,获得了最好的准确率、召回率和 F1 - 值。

综合以上实验及结果来看,本文设计的命名实体识别框架取得了预期的效果,其原因可归结于以下因素:一是构建了基于卷积循环神经网络的特征编码层,能够有效融合卷积和循环结构的优点来提升网络对序列数据局部空间特征和时序依赖特征的联合建模能力,使提取的特征对标注结果有更加显著、直接的影响;二是设计了基于改进 RNN 的标签解码层,充分利用 RNN 的时序连接结构来建模标签序列的上下文依赖关系,更有效地利用标签上下文信息来辅助预测;三是设计了分词与实体识别联合学习模式,将分词信息与实体信息纳入统一的标签模式下,结合相应的误差函数和模型训练算法来实现分词信息与实体信息的联合学习。

6　结论

本文研究了基于卷积循环神经网络的中文分词与实体识别联合学习框架,基于卷积循环神经网络构建特征编码层,实现对文本序列局部空间特征和长距离时序依赖特征的联合提取;基于改进的 RNN 构建了标签解码层,以建模标签序列内部的时序关联关系;同时设计中文分词和实体识别统一标注模式,实现了中文分词和实体识别的联合学习。在公开语料上的实验结果验证了该框架的有效性,尤其是对包含多个字的长实体的识别效果,更是取得了显著的提升,后续将研究把联合学习模型应用于特定领域的命名实体任务,以使该方法能够在更广阔的领域发挥作用。

参考文献(References)

[1] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329 - 340.
LIU Liu, WANG Dongbo. A review on named entity recognition[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(3): 329 - 340. (in Chinese)

[2] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proceedings of 18th International Conference on Machine Learning, 2001: 282 - 289.

[3] 俞鸿魁, 张华平, 刘群, 等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(2): 87 - 94.
YU Hongkui, ZHANG Huaping, LIU Qun, et al. Chinese named entity identification using cascaded hidden Markov model[J]. Journal on Communications, 2006, 27(2): 87 -

94. (in Chinese)

[4] 孙晓, 孙重远, 任福继. 基于深层条件随机场的生物医学命名实体识别[J]. 模式识别与人工智能, 2016, 29(11): 997 – 1008.
SUN Xiao, SUN Chongyuan, REN Fuji. Biomedical named entity recognition based on deep conditional random fields[J]. Pattern Recognition and Artificial Intelligence, 2016, 29(11): 997 – 1008. (in Chinese)

[5] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition [C]// Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 260 – 270.

[6] 张海楠, 伍大勇, 刘悦, 等. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报, 2017, 31(4): 28 – 35.
ZHANG Hainan, WU Dayong, LIU Yue, et al. Chinese named entity recognition based on deep neural network[J]. Journal of Chinese Information Processing, 2017, 31(4): 28 – 35. (in Chinese)

[7] 冯艳红, 于红, 孙庚, 等. 基于BLSTM的命名实体识别方法[J]. 计算机科学, 2018, 45(2): 261 – 268.
FENG Yanhong, YU Hong, SUN Geng, et al. Named entity recognition method based on BLSTM[J]. Computer Science, 2018, 45(2): 261 – 268. (in Chinese)

[8] ZHANG Y, YANG J. Chinese NER using lattice LSTM[J]. arXiv:1805.02023v2[cs.CL], 2018.

[9] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735 – 1780.

[10] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv:1508.01991v1[cs.CL], 2015.

[11] 李丽双, 郭元凯. 基于CNN-BLSTM-CRF模型的生物医学命名实体识别[J]. 中文信息学报, 2018, 32(1): 116 – 122.
LI Lishuang, GUO Yuankai. Biomedical named entity recognition with CNN-BLSTM-CRF[J]. Journal of Chinese Information Processing, 2018, 32(1): 116 – 122. (in Chinese)

[12] MA X Z, EDUARD H. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016: 1064 – 1074.

[13] JOZEFOWICZ R, ZAREMBA W, SUTSKEVER I. An empirical exploration of recurrent network architectures [C]// Proceedings of the 32nd International Conference on International Conference on Machine Learning, 2015, 37: 2342 – 2350.

[14] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673 – 2681.

[15] KINGMA D, BA J. Adam: a method for stochastic optimization [C]// Proceedings of the 3rd International Conference for Learning Representations, 2015.