

引用格式:程博,李卫红,童昊昕.基于 BiLSTM-CRF 的中文层级地址分词[J].地球信息科学学报,2019,21(8):1143-1151. [Cheng B, Li W H, Tong H X. Chinese address segmentation based on BiLSTM-CRF[J]. Journal of Geo-information Science, 2019,21(8):1143-1151.] DOI:10.12082/dqxxkx.2019.180654

基于 BiLSTM-CRF 的中文层级地址分词

程 博¹,李卫红^{1*},童昊昕²

1. 华南师范大学地理科学学院,广州 510631; 2. 航天精一(广东)信息科技有限公司,广州 510665

Chinese Address Segmentation based on BiLSTM-CRF

CHENG Bo¹, LI Weihong^{1*}, TONG Haoxin²

1. South China Normal University, Guangzhou 510631, China;

2. Guangdong Finest Planning Information Technology Limited Company, Guangzhou 510665, China

Abstract: Chinese word segmentation is a basic step in Chinese text processing and Chinese natural language processing. As a branch of Chinese word segmentation, Chinese address segmentation, which has become one of the hottest issues in Chinese word segmentation and geography research, is an important method to standardize Chinese address and conduct geocoding. Existing studies on Chinese address segmentation are mainly based on Statistical Machine Learning (SML) and Recurrent Neural Network (RNN). However, either of them cannot combine the advantages of the other. Furthermore, the current Chinese address segmentation methods lack address level segmentation and are too dependent on dictionaries and features. Therefore, this paper combined the four-word-position tagging set and Chinese hierarchical address characteristics to construct an address tagging system, and proposed a Chinese hierarchical address segmentation model (BiLSTM-CRF) which combines bidirectional long-term memory networks and conditional random fields algorithm. The proposed model utilizes the BiLSTM model to remember the characteristics of context address, while retaining the ability of the CRF algorithm to control the address tagging output by transferring probability matrix. In so doing, it has more powerful capability than the traditional statistical machine learning algorithms and RNN in the field of sequence labeling and word segmentation. To test the performance of BiLSTM-CRF on address samples marked by the address tagging system, CRF, LSTM, and BiLSTM were used to compare with BiLSTM-CRF and were respectively applied for training under the same condition as BiLSTM-CRF. We found that: (1) The segmentation effect of BiLSTM-CRF which is based on the Chinese address tagging system was better than the models for comparison, and the address tagging was more elaborate, in line with the actual address distribution. (2) The BiLSTM-CRF model had an accuracy of 93.4%, which was higher than the CRF (90.4%), LSTM (89.3%) and BiLSTM (91.2%) models. The overall address word segmentation performance and the effect of BiLSTM-CRF on each level address segmentation were more prominent than the other models. (3) The word segmentation performance of each model was correlated with the address level positively, i.e., the higher the address level, the better the word segmentation effect. The Chinese address tagging system and word

收稿日期:2018-12-13;修回日期:2019-04-29.

基金项目:广东省重大科技专项(2017B030305005)。 [**Foundation item:** Major Science and Technology Projects of Guangdong Province, No.2017B030305005.]

作者简介:程 博(1993-),男,湖北黄石人,硕士生,主要从事时空数据挖掘、自然语言处理。E-mail: 15625110283@163.com

*通讯作者:李卫红(1966-),女,四川泸州人,教授,硕士生导师,主要从事时空数据挖掘、犯罪地理研究。

E-mail: hongweili9981@163.com

segmentation model proposed in this study give a reference for the standardization of Chinese address, and provide the possibility to further improve the accuracy of geocoding technology. Future study can focus on fine-tuning the model to improve the model accuracy.

Key words: Chinese word segmentation; address tagging; hierarchical word segmentation; Long Short-Term Memory (LSTM); Bidirectional Long Short-Term Memory-Conditional Random Field (BiLSTM-CRF)

***Corresponding author:** LI Weihong, E-mail: hongweili9981@163.com

摘要:中文地址分词是中文地址标准化的基础工作和地理编码的重要手段,同时也是中文分词和地理研究领域关注的热点问题之一。针对当前中文地址分词方法缺乏地址层级切分和过多依赖词典和特征的问题,本研究结合四词位标注集和中文层级地址特点,构建针对中文层级地址分词的地址标注体系,并提出融合双向长短时记忆网络和条件随机场(BiLSTM-CRF)的中文层级地址分词模型。该模型既考虑了BiLSTM模型能够记忆上下文地址的特性,也保留了CRF算法可以通过转移概率矩阵控制地址标注输出的能力。针对该地址标注体系标注的训练地址样本,分别使用CRF、LSTM、BiLSTM与BiLSTM-CRF模型进行训练对比。结果表明:①基于中文地址标注体系的模型分词效果更佳,地址标注更为精细,符合实际地址分布情况;②BiLSTM-CRF模型精确度达到93.4%,高于CRF(90.4%)、LSTM(89.3%)和BiLSTM(91.2%),其整体地址分词性能和各层级地址分词效果相对于其他模型更突出;③各模型分词性能与地址层级保持一致,即地址层级越高,分词效果越好。本研究提出的中文地址标注体系和分词模型为开展中文地址标准化工作提供了方法参考,同时也为进一步提升地理编码技术的精准度提供了可能。

关键词:中文分词;地址标注;中文层级地址分词;长短时记忆网络(LSTM);双向长短时记忆和条件随机场模型(BiLSTM-CRF)

1 引言

中文分词是中文自然语言处理的基础,中文地址是描述空间地址信息的原始字符串,同一空间地址会有不同的描述形式。中文地址分词可以视为中文分词在地址编码领域的应用,它首先对连续的层级中文地址文本进行层级地址词组的切分,每个词组分别表示一级地址,之后将切分后的地址与已有标准空间地址库中的地址进行层级匹配获取地理坐标,从而实现地址文本和空间信息的关联,因此中文地址分词是中文地址标准化的基础工作^[1]。

因为中文语法体系存在词与词没有固定间隔,词的划分没有统一标准,“词”和“词组”的边界模糊^[2]等特征,所以精确地进行中文分词一直是中文分词方法研究的重点与难点。目前,国外众多学者已经开展相关中文分词方法的研究,主要包括基于词典^[3]、统计^[4]、深度学习^[5]、及上述方法相结合的混合分词方法^[6-7]。基于词典的分词方法具有构建简单、容易实现等优点,但存在未登录词和歧义识别以及命名体识别等中文分词难点无法解决的缺点。针对该方法适应性不好,扩展性差的问题,Xue^[8]、Tseng^[9]等先后提出了融合最大熵模型与四词位标注、条件随机场与字标注的统计分词方法。Zhao等^[10]在此基础上设计了结合6词位标注集与条件随机场模型的统计分词方法,这在一定程度上解决了

机械分词未登录词识别和切分歧义问题,同时避免了全切分方法因切分过多造成的分词效率低下问题^[11-12],但其分词效果受特征设定的限制,且存在特征过多容易出现模型训练过拟合的情况,因此需要降低人工标注以及特征工程过多带来的负面影响。而基于深度学习的分词方法极大地提高了中文分词的效率和计算性能,如针对未标注样本的监督学习问题,可以通过构建统一的神经网络架构和学习算法来降低人工标注的耗费^[13];针对中文分词和序列标注问题,可以借助深层神经网络来避免繁琐的特征工程,并且使用大规模的未标记数据来改善汉字内部表示,以最小的计算成本达到最优性能^[14]。为了能够进一步模拟中文分词中标签和上下文字符之间的复杂交互,Pei等^[15]提出一种新的Max-Margin张量神经网络(MMTNN)模型,并且此模型能够推广到序列标记等其他任务。而Chen等^[16-17]研究提出长短期记忆(LSTM)网络和门控递归神经网络(GRNN)进一步克服了局部上下文窗口大小的限制和梯度扩散问题,实现了中文分词的长时记忆能力,达到了更优的分词性能。但是上述研究较少涉及中文地址分词这一领域。

当前关于中文地址分词的研究主要集中于国内,在基于词典的地址分词方法上,已有学者在中文地址的数字表达方法^[18]、地理编码的匹配率和定位准确度^[19]、地名地址串的有效拆分^[20]等方面做出

卓有成效的研究成果。在基于统计的地址分词方面,主要依据统计决策树^[21]、地址语料库中字符串共现的统计规律^[22]来降低对城市地址词典的依赖程度,提高分词效率,同时也有基于深度学习的中文地址切分方法,进一步降低中文地址分词对词典的依赖,提高了切分正确率^[1,23]。但是国内在中文地址分词研究仍然存在2个较为突出的问题:①在中文地址分词任务的实现方面,研究主要基于字典和统计学习方法开展,仍然存在两种方法的固有弊端,而运用 LSTM 模型进行中文地址分词存在无法对地址标注输出进行限定的问题,从而产生标注转移错误;②以往研究大都采用四词位标注集,没有针对具体的分词任务设计具体的标注方法,仅使用四词位标注集能将地址正确切分为词组,却无法得知切分后词组对应的地址层级。

综上所述,中文地址分词除了具有普通中文分词的特点外,中文地址的层级特性及地址要素的随意性、多样性、歧义性等特性大大增加了中文地址分词的难度,而着力解决四词位标注集的地址层级和过多依赖词典和特征的问题是提升中文地址分词方法性能的关键。BiLSTM模型拥有强大的上下文记忆能力,不依赖词典、特征,可以解决未登录词和歧义问题,使用 GPU 可以极大提高训练速度,使用 BiLSTM 进行中文分词的效果更加显著^[24-26]。已有研究表明,BiLSTM 模型能够有效提升中文地址分词方法的性能^[23],而 CRF 算法可以通过转移概率矩阵控制地址标注输出。因此,本文基于中文地址的特点,利用四词位标注集设计了一套专用于中文地址分词的标注体系,基于该地址标注体系,将 BiLSTM 和 CRF 结合成为 BiLSTM-CRF 模型,实现中文地址的准确切分,同时将该模型与 CRF、LSTM 及 BiLSTM 模型进行训练对比,验证 BiLSTM-CRF 模型在中文层级地址分词上的优越性。

2 数据来源及研究方法

2.1 数据来源

本研究数据为航天精一(广东)信息科技有限公司提供的惠州市 151 000 条原始文本地址,每条地址数据记录包含不同的地址层级,地址层级皆从高到低。原始中文地址文本数据中存在重复地址、层级错乱、单条地址记录不完整、地址记录错误等问题。因此,对原始数据进行数据清洗,删除重复

记录,纠正层级错乱记录。记录不完整和错误记录则继续保留以增强模型的容错能力。经以上数据预处理后,剩下 15 万条地址文本。原始地址数据格式如表 1 所示。

表 1 原始地址数据格式

Tab. 1 Original address data format

样本编号	原始地址	地址层级关系
1	广东省惠州市仲恺区淡水街道办	省-市-区-街道
2	惠州市仲恺区沥林镇英光村民委员会	市-区-乡镇-村
3	广东省惠州市淡水爱民路基顿酒店	省-市-街道-街路巷-兴趣点
4	惠州市罗阳富力广场门口	市-兴趣点-方位

注 数据来源于广东省广州市航天精一(广东)信息科技有限公司。

根据 2009 年中国国家标准化管理委员会发布的数字城市地理信息公共平台地名/地址编码规则,并结合实际地址命名情况,可知中文地址大致分为行政区域地名、街巷名、小区名、门(楼)址和标志物名五级地址,各级地址包含相应的细分地址^[27]。地址数据需要经过标注后才能输入模型进行训练。多数研究使用(B,M,E,S)四词位标注集来对训练数据进行标注。本研究在四词位基础上增加了词位标注 A 来表示地址后缀(表 2),如“广东省”中“省”字既是单字词也是省级地址后缀,因此使用“省|S_PRO-A”来进行标注。各层级地址均根据所属的层级结合(B,M,E,S)标注集进行标注(表 3)。

表 2 词位标注集

Tab. 2 Word label set

词位	B	M	E	S	A
含义	词首	词中	词尾	单字词	后缀

2.2 研究方法

2.2.1 条件随机场

条件随机场(Conditional Random Field, CRF)是给定一组输入随机变量 X 条件下另一组输出随机变量 Y 的条件概率分布模型,其特点是假设输出随机变量 Y 构成马尔科夫随机场,通常可以使用线性链条件随机场解决中文地址标注问题。此时在条件概率模型 $P(Y|X)$ 中, Y 是输出变量,表示中文地址标注序列,也可称为状态序列; X 是输入变量,表示需要标注的中文地址序列(图 1)。在模型学习时,对训练数据集进行正则化的极大似然估计得到条件概率模型 $\hat{P}(Y|X)$;在预测时,对给定的输入地址序列,求出条件概率 $\hat{P}(Y|X)$ 最大的地址标注序列。

表3 层级地址标注体系
Tab. 3 Hierarchical address labeling system

地址要素类型	标注	含义	示例
省、自治区	B_PRO	省级地址词首	广 B_PRO
	M_PRO	省级地址词中	东 E_PRO
	E_PRO	省级地址词尾	省 S_PRO-A
	S_PRO-A	省级地址后缀单字词	
市、自治州、盟	B_CITY	市级地址词首	惠 B_CITY
	M_CITY	市级地址词中	州 E_CITY
	E_CITY	市级地址词尾	市 S_CITY-A
	S_CITY-A	市级地址后缀单字词	
区、县	B_COUNTY	区、县级地址词首	仲 B_COUNTY
	M_COUNTY	区、县级地址词中	恺 E_COUNTY
	E_COUNTY	区、县级地址词尾	区 S_COUNTY-A
	S_COUNTY-A	区、县级地址后缀单字词	
乡、镇、街道、派出所	B_STREET	乡镇级地址词首	淡 B_STREET
	M_STREET	乡镇级地址词中	水 E_STREET
	E_STREET	乡镇级地址词尾	街 B_STREET-A
	S_STREET -A	乡镇级地址后缀单字词	道 M_STREET-A
	B_STREET-A	乡镇级地址后缀词首	办 E_STREET-A
	M_STREET-A	乡镇级地址后缀词中	
	E_STREET-A	乡镇级地址后缀词尾	
村、社区	B_COM	村级地址词首	河 B_COM
	M_COM	村级地址词中	背 E_COM
	E_COM	村级地址词尾	村 B_COM-A
	S_COM-A	村级地址后缀单字词	民 M_COM-A
	B_COM-A	村级地址后缀词首	委 M_COM-A
	M_COM-A	村级地址后缀词中	员 M_COM-A
	E_COM-A	村级地址后缀词尾	会 E_COM-A
街、路、巷	B_ROAD	街路巷地址词首	先 B_ROAD
	M_ROAD	街路巷地址词中	烈 M_ROAD
	E_ROAD	街路巷地址词尾	东 E_ROAD
	S_ROAD-A	街路巷地址后缀单字词	路 S_ROAD-A
门楼牌	B_ML	门楼牌地址词首	20 B_ML
	M_ML	门楼牌地址词中	号 E_ML
	E_ML	门楼牌地址词尾	
建筑物	B_BUILD	建筑物地址词首	国 B_BUILD
	M_BUILD	建筑物地址词中	际 M_BUILD 大 M_BUILD
	E_BUILD	建筑物地址词尾	厦 E_BUILD
单元	B_UNIT	单元地址词首	12 B_UNIT
	M_UNIT	单元地址词中	单 M_UNIT
	E_UNIT	单元地址词尾	元 E_UNIT
房间	B_ROOM	房间地址词首	208 B_ROOM
	M_ROOM	房间地址词中	房 M_ROOM
	E_ROOM	房间地址词尾	间 E_ROOM
兴趣点	B_POI	POI地址词首	人 B_POI
	M_POI	POI地址词中	民 M_POI 医 M_POI
	E_POI	POI地址词尾	院 E_POI
方位	B_ORI	方位地址词首	附 B_ORI 近 E_ORI
	M_ORI	方位地址词中	前 B_ORI 方 E_ORI
	E_ORI	方位地址词尾	旁 S_ORI
	S_ORI	方位地址单字词	

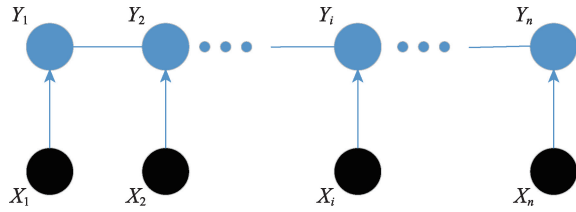


图1 条件随机场结构

Fig. 1 Conditional random field structure

条件随机场是对最大熵模型和隐马尔科夫模型的改进,解决了最大熵模型局部最优和标记偏置问题,避免了隐马尔科夫模型输出独立性假设无法考虑上下文特征以及无法进行特征选择的缺点。相对于最大熵模型和隐马模型,条件随机场在序列标注问题上的效果更好。但条件随机场需自定义特征,其模型分词效果受特征设定的限制,且存在特征过多容易出现模型训练过拟合和训练效率低下的问题。

2.2.2 LSTM和BiLSTM模型

循环神经网络(RNN)模型^[28]自提出以来,被大量运用在命名体识别、分词等自然语言处理领域,其具有记忆信息的能力,但经典RNN模型存在梯度消失和梯度爆炸等问题,难以解决长时记忆问题。长短时记忆网络(Long-Short Term Memory neural Network, LSTM)^[29]是在RNN基础上改进而来的一种神经网络模型,可以解决长期依赖问题。LSTM神经网络通过门结构来控制细胞状态中信息的增减,使用输入门(input gates)、忘记门(forget gates)和输出门(output gates)3种门结构以保持和更新细胞状态(图2)。由于LSTM的长时记忆能力,其在分词领域被广泛使用。

基于LSTM神经网络模型的中文地址分词是

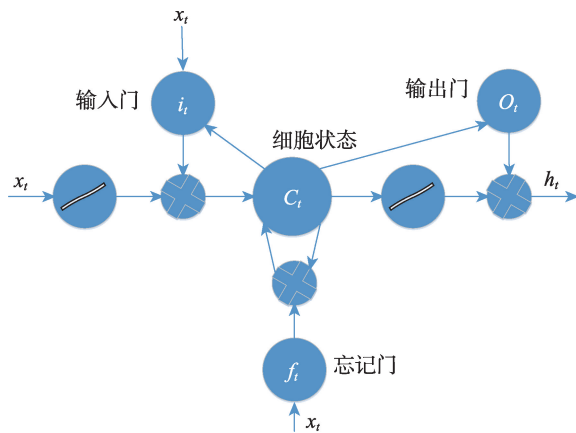


图2 LSTM神经元结构

Fig. 2 LSTM neuron structure

将中文地址序列标注任务看做一个多分类问题,其中输入特征 x 为中文地址序列,输出标签 y 为中文地址标注序列。模型一般需要3层:输入中文地址向量化层、LSTM网络层、Softmax标签推理层,如图3所示。

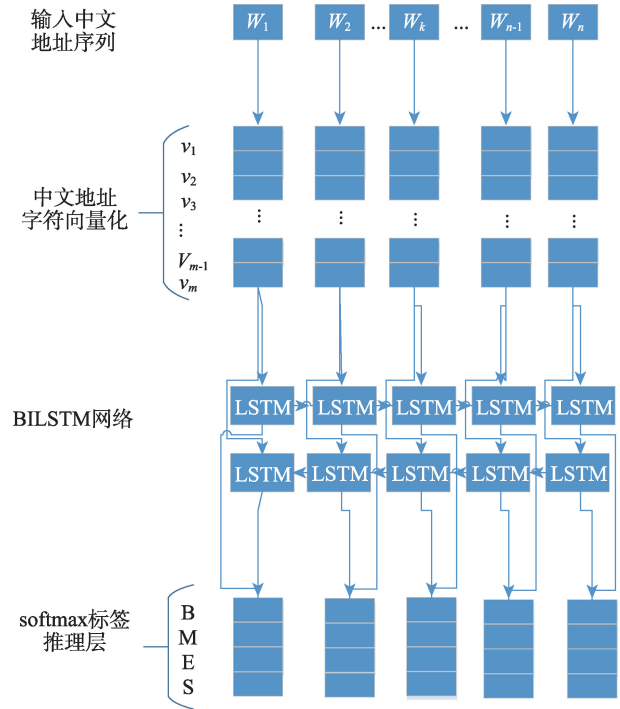


图3 基于BiLSTM的中文地址分词模型结构

Fig. 3 Chinese address segmentation model based on

BiLSTM structure

但是单向LSTM有其局限所在:只记忆过去的信息,无法考虑未来的上下文信息。因此便产生了BiLSTM (Bidirectional, Long-Short Term memory neural network, BiLSTM)神经网络,其具体思路来源于双向RNN网络模型。使用2个不同的LSTM神经网络层分别从中文地址的前端和后端进行遍历,这样便能保存2个方向的地址信息(图3)。相对于单向LSTM, BiLSTM既能保存前面的上下文地址信息,又能考虑未来的上下文地址信息,因此在中文地址分词任务中表现更佳。

2.2.3 BiLSTM-CRF模型

BiLSTM-CRF模型是将BiLSTM神经网络模型和条件随机场模型进行结合,即将原有BiLSTM模型中的Softmax层替换为条件随机场层。BiLSTM网络的输出为各个地址标注的概率,选择概率最大的标注则为对应的输出标注。但是BiLSTM仅仅选择概率最大的标注,而没有考虑输出标注间

的关系,即可能存在标注B后仍接标注B等类似情况。BiLSTM-CRF模型则将原有神经网络中的Softmax层替换为CRF层,通过将BiLSTM网络计算出的状态序列向量作为CRF层的输入,能够考虑标注之间的转移特征,即会考虑输出标注间的顺序,从而求出联合概率最大的地址标注序列。模型结构图4所示,该模型通过BiLSTM网络很好地记忆地址上下文的信息,并结合CRF层有效考虑了地址序列前后的标注。

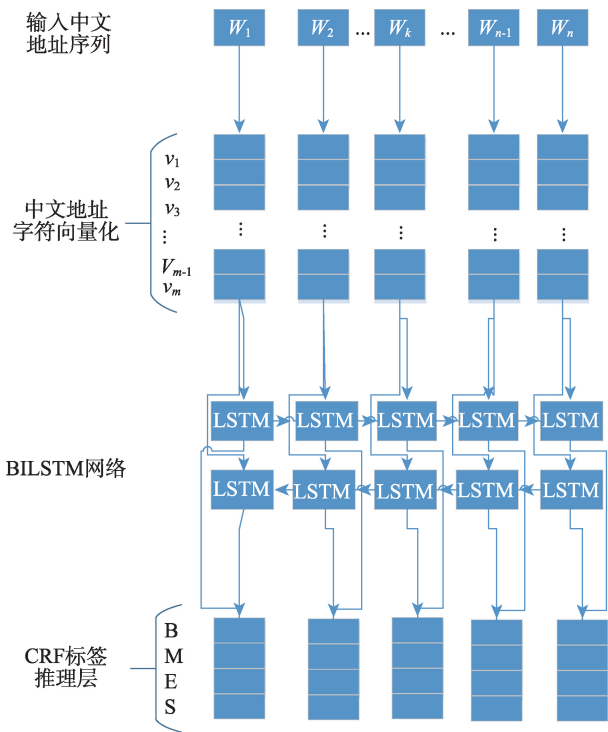


图4 基于BiLSTM-CRF的中文地址分词模型结构图

Fig. 4 Chinese address segmentation model based on BiLSTM-CRF structure

BiLSTM-CRF模型需要一个地址标注间的状态转移矩阵作为CRF层的输入参数,通过引入状态转移矩阵A,然后设定矩阵P为BiLSTM网络的输出,则观测地址序列X对应的地址标注序列 $y=(y_1, y_2, \dots, y_n)$ 的预测输出为:

$$s(X, y) = \sum_{i=1}^n (A_{y_i y_{i+1}} + P_{i, y_i}) \quad (1)$$

式中: A_{ij} 表示时序上从第*i*个状态转移到第*j*个状态的概率; $P_{i,j}$ 表示输入观察中第*i*个地址字符为第*j*个地址标注的概率; $s(X, y)$ 为计算出的地址标注序列得分,选择得分最大的序列即为最终的输出地址标注序列。利用维特比算法可以很好地计算出地址标注序列。

2.3 实验设置及模型评价标准

2.3.1 实验设置

本实验环境主要参数如表4所示。模型训练参数配置如表5所示,通过Dropout对数据进行正则化以降低测试误差,避免模型过拟合。模型初始学习率较大为0.1,通过Adam^[30]算法在训练过程中自适应调整。整个训练过程迭代40轮(epoch)。

表4 实验环境参数

Tab. 4 Parameters of the experimental environment

参数	值
CPU	Intel(R) Core(TM) i7-7700HQ@2.80GHZ 16G
GPU	NVIDIA GeForce GTX 1050 4G
操作系统	window 10 64 bits
编程语言	python
深度学习框架	Tensorflow

表5 模型训练参数

Tab. 5 Model training parameters

模型参数	值
词向量维度	200
批处理大小(batch_size)	128
神经网络隐藏单元数	100
Dropout率	0.5
神经网络层数	1
初始学习率	0.1
优化算法	Adam
epoch	40

2.3.2 模型评价指标

评价指标为精确率(Precision)和召回率(Recall)及调和均值F1-Measure,计算公式如下:

$$\frac{2}{F_1} = \frac{1}{P} + \frac{1}{R} \quad (2)$$

$$P = \frac{TP}{TP + FP} \quad (3)$$

$$R = \frac{TP}{TP + FN} \quad (4)$$

式中: F_1 为均值; P 为精确率; R 为召回率。

3 实验结果及分析

基于上述模型参数训练的4个中文地址分词模型在测试数据集上测试得到模型整体分词性能(表6)和各模型在各层级地址的分词性能(表7)。从表6可知,除LSTM模型3个指标和CRF模型召回率低于90%外,其余模型的相关指标均高于90%。

表6 模型整体分词性能

Tab. 6 Overall performances of the models in word segmentation

模型	精确率 P	召回率 R	均值 F_1
CRF	0.904	0.898	0.901
LSTM	0.893	0.887	0.890
BiLSTM	0.912	0.907	0.909
BiLSTM-CRF	0.934	0.929	0.931

表7 各级地址模型性能

Tab. 7 Address model performances at all levels

地址层级	均值 F_1			
	CRF	LSTM	BiLSTM	BiLSTM-CRF
省	0.977	0.968	0.986	0.993
市	0.971	0.961	0.983	0.997
区县	0.950	0.936	0.958	0.960
乡镇街道	0.936	0.925	0.932	0.942
村	0.915	0.906	0.914	0.922
街路巷	0.896	0.888	0.902	0.907
门楼牌	0.902	0.893	0.900	0.913
建筑物	0.897	0.879	0.911	0.918
单元	0.928	0.925	0.931	0.935
房间	0.918	0.910	0.915	0.919
兴趣点	0.873	0.857	0.877	0.905
方位	0.921	0.913	0.932	0.961

LSTM、CRF、BiLSTM 和 BiLSTM-CRF 这 4 个模型的指标值均依次递增。其中, BiLSTM-CRF 模型的 3 个指标均最高, 达到 93%, 模型整体分词性能最好, 其次为 BiLSTM 模型, 精确率高于 91%, 另外两个指标值也接近 91%。LSTM 模型的分词性能相对 CRF 模型和其他模型较差, 3 个指标值均低于 90%。CRF 模型召回率接近 90%, 精确率和 F_1 值高于 90%, 一定程度上优于 LSTM 模型。另外, 4 个模型的召回率相对于精确率较低, 原因可能是模型在训练数据上的效果比测试数据要好, 模型存在轻微的过拟合, 可以增加训练数据量或者修改 Dropout 率改善模型测试效果, 提高召回率。LSTM 模型的综合性能低于其他模型可能跟 LSTM 模型本身只能记忆过去信息, 无法获取未来信息的特点有关, 并且 LSTM 模型无法考虑输出标注间的关系, 这种缺点可能导致模型输出存在标注不连续的错误。CRF 模型的整体性能高于 LSTM 可能因为其能够设定大量的特征, 在充分拟合数据分布的情况下仍能考虑模型输出标注间的限制。而 BiLSTM 为双向记忆, 尽管没有限制输出标注间关系的能力, 但

通过记忆过去和未来地址信息的能力可以达到优于 CRF 模型的效果。最后, BiLSTM-CRF 模型综合了上述模型的优点, 既有强大的长期记忆能力, 也能考虑输出标注间转移特征, 因此各项指标均优于其他模型, 具备更佳的分词效果。

除了考察各模型的整体分词性能外, 本文也考察了各级地址的分词效果, 如表 7 展示了 F_1 综合指标表示的分词效果。从模型角度看, 各个地址层级中, BiLSTM-CRF 模型的分词效果仍然最佳, 其次为 BiLSTM 和 CRF, 均高于 LSTM 模型, 同模型整体分词性能保持一致。从由省至兴趣点的地址层级角度看, 整体上各模型分词性能呈下降趋势, 即整体上地址层级越高, 各模型分词性能越好, 地址层级越低模型分词性能逐渐下降。各级地址的分词性能中, 尤以省市两级为最, BiLSTM-CRF 模型的 F_1 指标值均超过 99%, 接近 100%。BiLSTM 和 CRF 模型的指标也分别超过 98% 和 97%, 较差的 LSTM 模型在省市级地址的识别上也有超过 96% 的性能。各模型在这两级地址上的优越表现可能是因为这两层地址的结构相对简单、种类较少, 而且均有明显的地址后缀, 更容易识别。区县、乡镇街道的分词性能紧随其后, 区县级地址在各个模型的分词性能大致与市级地址相差 1~2 个百分点, 而乡镇街道又与区县相差近 1~2 个百分点。导致这种性能差异的原因是区县级地址相对省市级地址较复杂多样, 但是相对于街道级地址却简单。街道相对于其他 3 个地址层级来说就相对繁杂, 除了地址种类较多外, 后缀“街道”也容易受地址层级“街路巷”中的“街”影响, 产生混淆, 因此分词效果相对欠佳。村、街路巷、门楼牌、建筑物、房间等地址层级的分词性能指标在 91% 附近徘徊, 这些层级地址种类繁多, 但是不同种类的样本却比较稀少, 此外结构复杂、后缀多或者没有后缀以及容易出现生僻地址等原因, 导致分词性能相对前几级地址较差。而单元和方位词两级地址结构简单, 分词效果较好, BiLSTM-CRF 的模型指标分别超过 93% 和 96%。各模型各层级地址中, POI 的分词效果最差, 除了结构复杂、种类繁多、没有明显地址后缀外, 最重要的原因是 POI 地址很容易与其他地址产生混淆, 如“惠州学院”一词, 可以使用 POI 表示, 也可以使用门楼牌表示, 使用后者表示时为“广东惠州市演达大道 46 号”。除此之外, 模型容易将其切分为“惠州”和“学院”两级地址, 前者为市级地址, 后者

可以为POI或者其他地址层级。并且,使用村一级地址也可以将“惠州学院”表示为“惠州学院社区”。此种情况,不胜枚举,造成POI的地址分词和识别相对其他地址层级效果较差,因此在处理模型训练数据集的时候也需要增加对POI地址的进一步处理以改进分词性能。

4 结论与讨论

已有研究没有根据各层级地址设计地址标注,且在地址分词研究上缺乏神经网络模型和统计机器学习模型的结合。本研究在四词位标注的基础上结合中文层级地址的特点设计了一套中文地址标注体系,并将BiLSTM网络和CRF模型相结合,充分利用二者的优势,然后使用根据该标注体系标注后的数据来训练CRF、LSTM、BiLSTM和BiLSTM-CRF模型并加以对比,实验结果表明:

(1)该中文地址标注体系相对于(B, M, E, S)四词位标注集更加符合实际地址分布状况,地址标注更具体,粒度更细致,基于该标注体系训练的中文地址分词模型分词效果更佳,也更容易为后续中文地址标准化工作所利用。

(2)神经网络模型和统计机器学习模型相结合在地址分词上的效果优于其中任一单一模型效果。BiLSTM-CRF模型的整体分词性能和各层级地址分词效果优于其他3个模型,其次为BiLSTM和CRF模型,LSTM模型分词效果相对较差。

(3)各模型在各层级地址的分词性能是与地址的层级基本保持一致,即层级越高,分词效果越好。其中,省、市、区县三级地址分词性能远优于其它层级,乡镇街道、单元、方位词比村、街路巷、门楼牌、房间表现较好。POI尤为值得关注,其分词性能差于其它地址层级。

因此,后续研究可从以下3个方面进行深入探讨:①六词位标注集和四词位标注集在中文地址分词中的对比;②增加模型层数和隐藏层神经元数量能否提升中文地址分词效果;③不同的Dropout率、学习率和词向量维度等参数设置对中文地址分词的性能影响,以及引入注意力机制的影响。

参考文献(References):

[1] 张文豪,卢山,程光.基于LSTM网络的中文地址分词法的设计与实现[J].计算机应用研究,2018,35(12):1-2.
[Zhang W H, Lu S, Cheng G. Design and implementa-

tion of chinese address segmentation method based on LSTM networks[J]. Application Research of Computers, 2018,35(12):1-2.]

- [2] 郭正斌,张仰森.基于定长序列的双向LSTM分词优化方法[J].郑州大学学报(理学版),2018,50(2):97-101. [Guo Z B, Zhang Y S. A optimization method of bidirectional LSTM based on fixed length sequence for word segmentation[J]. Journal of Zhengzhou University (Natural Science Edition), 2018,50(2):97-101.]
- [3] 莫建文,郑阳,首照宇,等.改进的基于词典的中文分词方法[J].计算机工程与设计,2013,34(5):1802-1807. [Mo J W, Zheng Y, Shou Z Y. Improved chinese word segmentation method based on dictionary[J]. Computer Engineering and Design, 2013,34(5):1802-1807.]
- [4] 褚颖娜,廖敏,宋继华.一种基于统计的分词标注一体化方法[J].计算机系统应用,2009,18(12):55-58. [Chu Y N, Liao M, Song J H. Integrated chinese words segmentation and labeling based on Statistic Method[J]. Computer Systems & Applications, 2009,18(12):55-58.]
- [5] Che J L, Tang L W, Deng S J, et al. Chinese word segmentation based on bidirectional GRU-CRF model[J]. International Journal of Performability Engineering, 2018,14(12):3066-3075.
- [6] 蒋建洪,赵嵩正,罗玫.词典与统计方法结合的中文分词模型研究及应用[J].计算机工程与设计,2012,33(1):387-391. [Jiang J H, Zhao H Z, Luo M. Analysis and application of chinese word segmentation model which consist of dictionary and statistics method[J]. China Computer Federation Magazine, 2012,33(1):387-391.]
- [7] Dong C H, Zhang J J, Zong C Q, et al. Character-based LSTM-CRF with radical-level features for chinese named entity recognition[C]. Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016 (10102):239-250.
- [8] Xue N W. Chinese word segmentation as character tagging[J]. Computational Linguistics and Chinese Language Processing, 2003,8(1):29-48.
- [9] Tseng H, Chang P, Andrew G, et al. A conditional random field word segmenter for sighan bakeoff 2005[C]. Proceeding of the Fourth SIGHAN Workshop on Chinese Language Processing, Jeju Island, Korea, 2005:168-171.
- [10] Zhao H, Huang C N, Li M. An improved chinese word segmentation system with conditional random field[C]. Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing, Sydney, Australia, 2006:162-165.
- [11] Kneser R, Ney H. Improved backing-off for m-gram language modeling[C]. 1995 International Conference on Acoustics, Speech, and Signal Processing, Michigan,

- USA, 1995.
- [12] Lillenberg J, Zhu Y, Zhang Y. Support vector machines and Word2vec for text classification with semantic features [C]. 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC). IEEE, 2015.
- [13] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011,12(8):2493-2537.
- [14] Zheng X, Chen H, Xu T. Deep learning for Chinese word segmentation and POS tagging[C]. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Washington, USA, 2013:647-657.
- [15] Pei W, Ge T, Chang B. Max-margin tensor neural network for chinese word segmentation[C]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, 2014:293-303.
- [16] Chen X, Qiu X, Zhu C, et al. Long short-term memory neural networks for chinese word segmentation[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 2015: 1197-1206.
- [17] Chen X, Qiu X, Zhu C, et al. Gated recursive neural network for chinese word segmentation[C]. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Beijing, China, 2015,1:1744-1753.
- [18] 张雪英, 闫国年, 李伯秋, 等. 基于规则的中文地址要素解析方法[J]. 地球信息科学学报, 2010,12(1):9-16. [Zhang X Y, Lv G N, Li B Q, Chen W J. Rule-based approach to semantic resolution of chinese addresses[J]. Journal of Geo-information Science, 2010,12(1):9-16.]
- [19] 余靖毅, 邬伦, 高勇. 基于 Storm 的地理编码引擎[J]. 地球信息科学学报, 2015,17(12):1431-1441. [Yu J Y, Wu L, Gao Y. A geocoding engine based on storm[J]. Journal of Geo-information Science, 2015,17(12):1431-1441.]
- [20] 赵阳阳, 王亮, 仇阿根. 地址要素识别机制的地名地址分词算法 [J]. 测绘科学, 2013,38(5):74-76. [Zhao Y Y, Wang L, Qiu A G. An improved algorithm for address segmentation[J]. Science of Surveying and Mapping, 2013,38(5):74-76.]
- [21] 应申, 李威阳, 贺彪, 等. 统计决策树下的城市地址集中文分词[J]. 武汉大学学报·信息科学版, 2018(12):1-9. [Ying S, Li W Y, He B, et al. Chinese segmentation of city address set based on the statistical decision tree[J]. Geomatics and Information Science of Wuhan University, 2018(12):1-9.]
- [22] 谢婷婷, 严柯. 基于统计的中文地址位置语义解析方法研究[J]. 软件导刊, 2017,16(10):19-21. [Xie T T, Yan K. The method of semantic resolution of chinese addresses based on statistics[J]. Software Guide, 2017,16(10):19-21.]
- [23] 李一, 刘纪平, 罗安. 深度学习的中文地址切分算法[J]. 测绘科学, 2018,43(10):107-111. [Li Y, Liu J P, Luo A. Chinese address segmentation algorithm based on depth learning[J]. Science of Surveying and Mapping, 2018,43(10):107-111.]
- [24] Yao Y, Huang Z. Bi-directional LSTM recurrent neural network for chinese word segmentation[C]. International Conference on Neural Information Processing, Kyoto, Japan, 2016:345-353.
- [25] Zhang M, Yu N, Fu G. A simple and effective neural model for Joint word segmentation and POS tagging[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018,26(9):1528-1538.
- [26] 金宸, 李维华, 姬晨, 等. 基于双向 LSTM 神经网络模型的中文分词[J]. 中文信息学报, 2018,32(2):29-37. [Jin C, Li W H, Ji C, et al. Bi-directional long short-term memory neural networks for chinese word segmentation[J]. Journal of Chinese Information Processing, 2018,32(2):29-37.]
- [27] GB/T 23075-2009 数字城市地理信息公共平台地名/地址编码规则. [GB/T 23075-2009 The rules of coding for address in the common platform for geospatial information service of digital city.]
- [28] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Nature, 1986, 323(6088):533.
- [29] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural computation, 1997,9(8):1735-1780.
- [30] Kingma D P, Ba J. Adam: A method for stochastic optimization[C]. International Conference on Learning Representations (ICLR), San Diego, United States, 2015,3:1-13.