一种基于IDLSTM+CRF的中文主地域抽取方法

童逸琦, 叶培根, 付彪, 陈毅东*, 史晓东
厦门大学信息学院人工智能系, 福建, 厦门, 361005
{yqtong，pgye，biaofu}@stu.xmu.edu.cn，{ydchen，mandel}@xmu.edu.cn

摘要

新闻文本通常会涉及多个地域，主地域则描述了文本舆情内容的地域属性，是进行舆情分析的关键属性。目前深度学习领域针对主地域自动抽取的研究还比较少。基于此，本文构建了一个基于IDLSTM+CRF的主地域抽取系统。该系统通过地名识别、主地域抽取、主地域补全三大模块实现对主地域标签的自动抽取和补全。在公开数据集上的实验结果表明，我们的方法在地名识别任务上要优于BiLSTM+CRF等模型。而对于主地域抽取任务，目前还没有标准的中文主地域评测集合。针对该问题，我们标注并开源了1226条验证集和1500条测试集。最终，我们的主地域抽取系统在两个集合上分别取得了91.7%和84.8%的抽取准确率，并成功运用于线上生产环境。

关键词： 舆情分析 ； 主地域抽取 ； IDLSTM+CRF

A Chinese Main Location Extraction Method based on IDLSTM+CRF

Yiqi Tong, Peigen Ye, Biao Fu, Yidong Chen*, Xiaodong Shi
School of Informatics, Xiamen University, Fujian, Xiamen, 361005
{yqtong，pgye，biaofu}@stu.xmu.edu.cn，{ydchen，mandel}@xmu.edu.cn

Abstract

The main location describes the regional attributes of the public opinion content of the text, which is a key attribute for public opinion analysis. However, there are still relative few researches on the Chinese main location extraction. In this paper, we propose an IDLSTM+CRF based main location extraction system, which contains three parts: location recognition, main location extraction and main location completion. The experimental results showed our model is better than mainstream models such as BiLSTM+CRF in location entity extraction. Furthermore, there is currently no standard Chinese main location evaluation sets. To solve this problem, we annotated a dev set and a test set, which contains 1226 and 1500 articles, respectively. Finally, our model can achieved 91.7% and 84.8% accuracy on these two sets and successfully applied to the online environment.

Keywords: public opinion analysis , main location extraction , IDLSTM+CRF

*通信作者。

1 引言

随着互联网文本的指数式增长，人们越来越面临信息爆炸问题。如何帮助人们高效地从非结构化数据(Unstructured data)中获取关键知识成为重要的需求，也成为了信息抽取(Information extraction, IE)和信息检索(Information retrieval, IR)等领域的热门研究课题。近年来，人们普遍关注突发事件新闻(刘佳琪and 罗永莲, 2019)，而新闻文本的地域属性蕴含了新闻事件发生的地点，是对新闻事件进行统计、分析的重要参考维度，因此研究如何从事件新闻中识别地名关键词具有重要的意义。有研究表明(马雷雷et al., 2016)，至少70%的文本文档包含以地名形式表达的地理位置参考信息。而这些地名中往往只有一个或两个是与文本描述的事件相关联的，我们将这种能够描述事件地理位置信息的核心地名定义为主地域。目前国内开展主地域抽取的工作还比较少，本文研究的核心就是从非结构化文本中自动的抽取主地域并对其进行补全以推动舆情分析、文本摘要、推荐系统等下游任务的发展。

具体的，本文对非结构化文本的主地域属性解析采用了流水线(Pipeline)的形式，其包含三个关键步骤：地名识别，主地域抽取，主地域补全。地名识别属于命名实体识别(Named entity recognition, NER)领域。得益于计算资源和训练数据的增长，结合条件随机场(Conditional random field, CRF) (Peng and Dredze, 2015)和深度学习(Deep learning)的模型如循环神经网络(Recurrent Neural Networks, RNN) (黄炜et al., 2019)、卷积神经网络(Convolutional Neural Networks, CNN) (曹春萍and 关鹏举, 2019)等取得了较好的成绩。其中，最典型的代表分别是双向长短时记忆网络(Bidirectional Long-Short Term Memory, BiLSTM) (Huang et al., 2015)和膨胀卷积神经网络(Iterated Dilated CNN, IDCNN) (Strubell et al., 2017)。为了结合BiLSTM和IDCNN的特征抽取优点，本文在两者的基础上，提出了一个基于IDLSTM+CRF和注意力机制的地名实体识别模型，在公开的命名实体识别数据集上的实验结果表明，该模型相较前二者取得了更好的成绩。

针对主地域抽取过程中出现的同一地名有多种表述的关键问题，本文首先对抽取出的地名实体进行同一化操作，再利用地名实体在文本中的位置信息、词频信息等来构建特征向量，进而通过求解主特征向量来计算各实体的主地域权重，最终获得不同地名实体在文本中的重要程度。中国地幅辽阔，而中文表述又通常存在缺省的现象，为了让读者或研究人员能清晰的将缺省地名与实际地理位置产生直观映射。本文根据中国的行政区划，构建了省/市/县三级的地域知识库，并参考分词中的最大匹配算法(Maximum matching algorithm)来递归的实现地域补全。例如“思明”，通过补全算法可以将其补全为“福建省厦门市思明区”。

数据集对深度学习的贡献是巨大的，如斯坦福大学开源的阅读理解数据集SQuAD (Rajpurkar et al., 2016; Rajpurkar et al., 2018)，极大的推动了机器阅读理解(Machine reading comprehension, MRC) 和问答(Question answering, QA) 系统的发展。因为国内对主地域识别的研究较少且分散，目前还没有统一的评测集合。为解决该问题，本文构建两个主地域数据集，共包含2726条新闻和微博数据，为该领域进一步的工作和相关实验提供了基础条件。我们的数据和代码开源于`https://github.com/zgzjdx/Chinese_Main_Location_Extraction`。

2 相关工作

2.1 地名识别

传统的地名识别方法主要有两种：a).基于规则的地名提取方法(张雪英et al., 2010; 赵英et al., 2017)，因为我国的地域描述包含“专用名+通用名”的构词规则，如“江苏省”，“江苏”为专用名，“省”为通用名。该方法利用上述构词规则构建规则库来对文本进行匹配，从而实现地名提取。虽然该方法的识别精确率较高，但缺点较为明显，中文文本尤其是互联网文本的书写较为随意，存在着大量通用名缺省的现象，导致其识别的召回率不高。b).基于知识库的地域提取方法(刘瑜et al., 2007; 唐旭日et al., 2010; 马雷雷et al., 2015)，该方法通过收集我国的行政区划或根据地名本体(Ontology)来构建地域知识库，进而通过文本与知识库匹配实现地名识别。该方法费时、费力，且需要大量专家知识来构建并维护地域知识库。

目前主流的地名识别方法有如下两类：a).机器学习(Machine learning)方法(张杰et al., 2008; 俞鸿魁et al., 2006; Maimaiti et al., 2017)，这类方法的主流模型有最大熵模型(Maximum entropy model, ME)、隐马尔可夫模型(Hidden markov model, HMM)、条件随机场模型等。其中，条件随机场模型效果最好，使用最广泛，因为其考虑了时间序列，构成的无向图(Undirected graph)模型天然与文字序列相吻合。跟传统地名识别方法比，机器学习方法

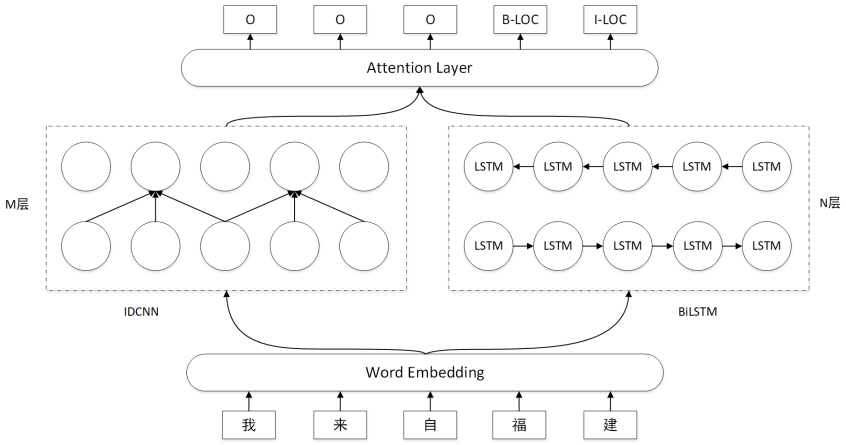

图 1. IDLSTM+CRF模型整体架构

效果更好，一些研究结果(孙镇and 王惠临, 2010)表明该方法能在封闭测试的环境下取得90%以上的识别准确率，但该方法需要训练数据和人工构建特征模板。b).深度学习方法(武惠et al., 2019; 李妮et al., 2020)，该方法和机器学习最大的区别在于采用了深度神经网络(Deep neural networks, DNNs)来自动学习和抽取特征，从而提升模型的泛化能力，使其在开放测试的环境下也能取得较好的效果。在基于深度学习的地名识别领域中，最主流的模型为BiLSTM+CRF，其由一个从前端到后端的LSTM和后端到前端的LSTM构成。LSTM(Long Short Term Memory, 长短期记忆网络)是在RNN的基础上改进而来，用以解决梯度爆炸问题(Gradient exploding problem)，其使用输入门(Input gate)、忘记门(Forget Gate)以及输出门(Output gate)3种门结构来保持和更新状态。因为BiLSTM每一个时间步的计算依赖于上一个时间步的计算结果，导致其计算速度一般。为了解决该问题，研究人员又提出IDCNN+CRF模型，该模型通过并行的空洞卷积、池化操作来提升运算速度，在和BiLSTM+CRF模型保持相仿识别性能的同时，该模型在训练和推理上有明显优势。

2.2 主地域抽取

随着地名识别技术的发展，也有部分研究人员开始对文本中的热度地名识别和抽取开展深入的探索。石桢and 姚天 (2013)等人提出了基于统计和规则的针对地名文化类文本的核心地名抽取方法，在100篇新闻类文本上取得81%的抽取准确率。李照航et al. (2015)提出了基于ATF-PDF(Average Term Frequen-cy-Proportional Document Frequency)模型的词汇权重方法，该方法是TF-IDF(Term Frequen-cy-Proportional Document Frequency)的一种变种，用于计算热度地名在旅游文本中的综合权重。智烈慧et al. (2016)，在ATF-PDF的基础上，引入高热度地名的共现频次矩阵，实现众包旅游文本的热度地名挖掘，并将计算结果以共现矩阵和三元组共同存储的方式呈现。钟翔et al. (2016)认为ATF-PDF模型仅考虑了文本中地名的词频等统计信息，而忽略了文本中地名间的共现传递关系，进而提出了一种基于链接分析的网页文本核心地名提取方法，该方法通过PageRank算法来计算文本中各个地名在共现网络中的链接权重，从而实现对具有显著焦点特征或导航枢纽特征的重要地名抽取。舒时立et al. (2019)提出了基于隶属关系的地名树结构的最佳空间尺度新闻事件地点抽取方法，该方法通过引入虚父节点和标准地名知识库来构建隶属关系地名树，结合最小包围盒(Minimum bounding box)算法实现最佳空间尺度的选择，最终通过地名实体权重实现事件发生地的排序及选取。

3 基于IDLSTM+CRF的主地域识别

3.1 地名识别模块

地名识别是主地域抽取的前提和基础，为了动态的结合两类主流特征提取器的优点，本文提出了一个基于IDLSTM+CRF的地名识别模型，结合注意力机制来学习和分配两个网络的权重。如图1所示，该模型由BiLSTM特征提取器、IDCNN特征提取器、一个可选的注意力机制和条件随机场组成。

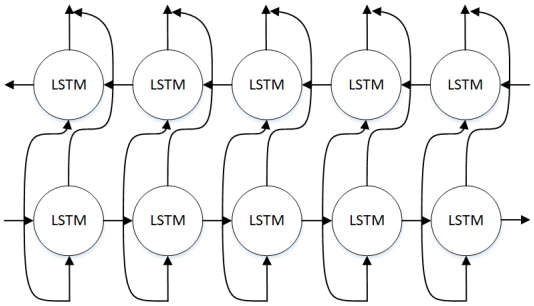

图 2. BiLSTM模型

3.1.1 BiLSTM层

为了捕获文本的时间序列特征，我们使用BiLSTM网络对输入序列进行特征提取，其内部结构如图2所示。对于输入序列S，其通常由若干个$token$组成，即$S = \{t_1, ..., t_m\}$，其中m代表输入序列的长度。本文通过在维基百科上预训练的词向量将其转化为语义特征序列，此时$S = \{x_1, ..., x_m\}$，其中x_i表示t_i的词嵌入表示。我们分别将其输入双向LSTM中，获得对应的隐状态特征。

对于BiLSTM，在t时刻，我们进行如下的计算来得到其隐状态表示。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{1}$$
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{2}$$
$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{3}$$
$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4}$$

其中，f_t，i_t,o_t分别表示长短期记忆网络的遗忘门，输入门和输出门在t时刻的输出，$\tilde{C}_t$代表t时刻的细胞状态，σ代表sigmoid激活函数，h_{t-1}代表$t-1$时刻的隐状态，$\cdot$代表点积，W和b分别为权重矩阵和偏置项。最终，根据遗忘门和输入门进行细胞状态更新，再结合输出门和$tanh$激活函数得到其双向的隐状态表示。

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{5}$$
$$h_t = o_t * tanha(C_t) \tag{6}$$
$$h_t = h_{forward_t} + h_{backward_t} \tag{7}$$

其中，$h_{forward_t}$和$h_{backward_t}$分别代表t时刻前向LSTM的隐状态和后向LSTM的隐状态，将二者进行拼接后得到最终的隐状态表示h_t。

3.1.2 IDCNN层

为了捕获文本的局部特征，本文使用IDCNN进行二次特征提取。在原始的卷积神经网络中，其卷积核是连续滑动的，而IDCNN在卷积操作上增加了一个膨胀宽度参数来增大感受野(Receptive field)，即进行卷积操作时会跳过中间空洞区域，从而可以减少网络层数来避免过深的层数导致模型产生过拟合(Overfitting)问题。图3为IDCNN的模型结构图，其膨胀步长设置为4，每一步膨胀卷积的过滤宽度为3。

具体的，我们对每一个x_t进行如下的计算来得到其隐状态表示。

$$c_t = W_t \oplus_{k=0}^{r} x_{t \pm k\delta} + b_t \tag{8}$$

其中，$\oplus$代表向量拼接操作，r代表过滤宽度，δ代表膨胀步长，当取其值为1时，便是标准的卷积操作，W和b分别代表权重矩阵和偏置项。

3.1.3 注意力层

为了更好的融合BiLSTM和IDCNN的抽取特征，本文设计了一个基于注意力机制的网络层来学习序列与序列之间的联系，从而动态的分配权重。注意力机制首先在计算机视觉领

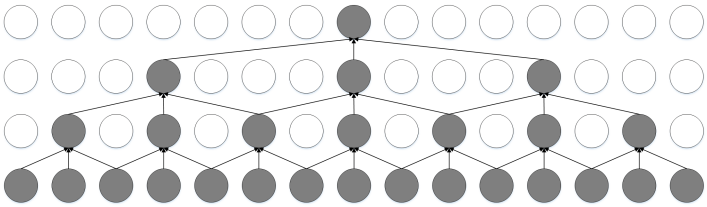

图 3. IDCNN模型

域得到应用(Mnih et al., 2014)，随后Bahdanau et al. (2014)将其引用于机器翻译(Machine translation)任务上并取得了巨大的突破，引发了一波研究注意力机制的热潮。

注意力机制事实上是对输入权重分配的关注，具体的，对于BiLSTM的隐状态表示h_t和c_t，本文根据下列公式进行最终隐状态表示H_t的计算。

$$c_t = W \cdot c_t + b \tag{9}$$

$$\partial = softmax(V^t tanh(\frac{Q \cdot h_t + U \cdot c_t}{\sqrt{dim}})) \tag{10}$$

$$H_t = tanh(\partial \cdot h_t + (1 - \partial) \cdot c_t) \tag{11}$$

其中，W和b是维度变换矩阵和偏置项，Q、U和V为注意力层的权重矩阵,其三者的维度均为dim，并最终通过$softmax$操作求解其权重。

3.1.4 CRF层

因为在地名识别任务中，一个token的标签与其相邻token的标签存在一定的联系，例如I-LOC不会出现在B-ORG的后面。因此，我们利用命名实体领域常用的条件随机场进行标签概率计算。该模型会计算输入序列的最优联合概率，即全局最优的标注序列。设输入序列S的标签序列Y为$\{y_1, \ldots, y_m\}$，本文根据下列公式进行评估分数$S(x, y)$的计算。

$$S\left(x,\ y\right) = \sum_{i=1}^{m} M_{y_i, y_{i+1}} + \sum_{j=i}^{m} P_{x_j, y_j} \tag{12}$$

其中，M为转移矩阵，$M_{y_i,y_{i+1}}$代表从y_i标签转移到y_{i+1}标签的概率。P_{x_j,y_j}代表第j个字被标记为y_j的概率，m为输入序列长度。

最终在训练过程中，采用极大似然估计原理对其进行优化，其计算如下:

$$log(P\left(y|x\right)) = \ S\left(x,\ y\right) - \sum_{i=1}^{m} S\left(x,\ y_i\right) \tag{13}$$

3.2 主地域抽取模块

本模块依赖实体的统计信息进行主地域抽取。但对于地名识别模块识别出的实体，模型并不能分辨同一实体的不同表述形式，如“厦门”和“厦门市”代表的是同一地理位置信息“福建省厦门市”。为解决该问题，我们构建了一个同一地名库(公布开源地址)，将表示同一地理位置的不同表述实体映射到统一的表述上。

设输入文本为d，地名识别模块识别出的实体集合为$\{x_1, ..., x_m\}$，经同一地名处理后的实体集合为$\{x_1, ..., x_n\}_{n<m}$，不同的实体代表不同的地理位置信息。我们统计其长度、词频和位置信息，并将其转化为特征因子。

对于长度和词频信息，我们采用简单的非线性函数进行转化:

$$fre_{w_i} = \frac{f_{w_i}}{f_{w_i} + 1},\ len_{w_i} = \frac{l_{w_i}}{l_{w_i} + 1} \tag{14}$$

其中，f_{w_i}和l_{w_i}为实体w_i的词频和长度，fre_{w_i}和len_{w_i}分别为实体w_i的词频特征因子和长度特征因子。

文本类型	位置	位置权重
新闻长文本	标题	5
	文首	3
	正文	1
	文末	3
微博短文本	文首	5
	正文	1
	文末	3

表 1. 实体的位置权重转化规则.

对于实体的位置信息，我们首先根据先验知识认为出现在标题的地名实体最能反映主地域信息，而文首和文末出现的地名实体要比正文中的地名实体更具有价值，因此本文根据表1所示的规则将实体的位置信息转化为位置权重。其中，对于微博短文本和新闻长文本，我们设置不同的权重规则。对于长文本，我们设前15%的正文为文首，后15%的正文为文末。对于短文本，我们设前30%的正文为文首，后30%的正文为文末。

而对于w_i的位置权重θ_i，本文通过成对比较法和主特征向量求解将其转化为位置特征因子loc_{w_i}。成对比较法是一个两两相比较构成的判断矩阵，对于位置权重集合$\{\theta_1, \ldots, \theta_m\}$，我们根据成对比较法构建下列的特征方阵$A_{m\times m}$。

$$A_{m\times m} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \cdots & \cdots & \cdots \\ a_{m1} & \cdots & a_{mm} \end{bmatrix}, \ a_{ij} = \frac{\theta_i}{\theta_j}, \ i, j \leq m \tag{15}$$

根据石桢and 姚天 (2013)的研究，当特征矩阵A为一致性矩阵时，即$a_{ij} \cdot a_{jk} = a_{ik}$时，矩阵的主特征向量就是实体的位置特征因子的近似度。主特征向量是指主特征值对应的向量，而主特征值是指模最大的特征值。因此我们使用幂迭代法对该矩阵求解主特征值δ_{max}和主特征向量$W_{m\times 1}$，其第i维的值对应的便是w_i的位置特征因子。

我们按照下列公式计算地名实体的最终位置权重:

$$Score_w = softmax(a * fre_w + b * len_w + c * loc_w + d) \tag{16}$$

$$softmax = \frac{e^{score_w i}}{\sum_{j=1}^{m} e^{score_{wj}}} \tag{17}$$

其中a，b和c为模型参数，d为偏置项。在归一化函数之后，我们得到各个地名实体的主地域分数，并且排序，最终取N-best作为主地域抽取结果[1]。

3.3 主地域补全模块

为了使抽取结果更加直观的展现地理位置信息，本小节实现了一个基于最大匹配算法的主地域补全模块。考虑到地域补全离不开外部知识的补充，本文从中华人民共和国民政局全国区划查询平台[2]爬取并构建了省/市/县三层架构的地域知识库[3]，并以JSON格式保存。

主地域补全域算法如算法1所示，我们设最小匹配长度为2，其中$longest_match$函数展现的是正向最大匹配过程，$split$函数是递归的入口，其递归的将地名解析为省/市/县三个字段。经过主地域补全模块，可以将缺省的地名实体如“恩施”，补全为“湖北省恩施土家苗族自治州”。

4 实验结果和分析

4.1 数据集

本文的地名识别模型采用公开的中文命名实体识别数据集bakeoff2006进行训练[4]，该数据集

[1]我们分别实验了阈值限制和N-best两种策略来筛选主地域，结果表明N-best的效果要好于阈值限制。
[2]http://xzqh.mca.gov.cn/map
[3]https://github.com/zgzjdx/Administrative-Division-Of-China
[4]https://github.com/zgzjdx/BERT-NER/tree/master/data

算法 1 基于正向最大匹配算法的地域补全

输入: 待匹配字符串str，省/市/县三级地域知识库lexicon
输出: 补全结果*address*

```
def longest_match(string, sub_lexicon):
    target, index = " ", 0
    for i in range(1, len(string) do
        for item in sub_lexicon do
            if string[:i+1] in item then
                index, target = i + 1, item
            end if
        end for
    end for
return index, target

def split(str, lexicon, level=1):
    (index, target) = longest_match(str, lexicon)
    if match successfully then
        split(str[index:], lexicon[target], level+1)
        address[level] = target
    elif fail to match at this level then
        for (target, sub_lexicon) in lexicon.items() do
            if split(str, sub_lexicon, level+1) then
                address[level] = target
            end if
        end for
    else
    # the end of the match
        return False
```

集合	句子数	地名	组织机构名	实体总计
训练集	20864	39102	46966	86068
验证集	2318	4532	4929	9461
测试集	4636	8606	10941	19547

表 2. 命名实体识别数据集的统计情况.

以BIO的形式标注了地名、组织机构名、人名三类实体信息，并划分好了训练集、验证集、测试集。考虑到组织机构名能对主地域的识别起辅助判断作用，且为了避免人名实体对地名识别的干扰，本文剔除了人名实体，仅保留地名、组织机构名两类实体。剔除人名实体后数据集的统计信息如表2所示。

为了衡量模型的主地域抽取效果，我们以双工的形式标注了一份主地域验证集和一份主地域测试集，分别包含1226条和1500条文本，并对测试集标签进行了人工补全，文本内容包含来源于新闻、微博等。两份主地域数据集的一些统计信息如表3所示。从表格数据可以得到，验证集平均每条文本包含3.98个主地域标签，测试集平均每条文本包含1.12个主地域标签。因为验证集中以新闻类长文本为主，测试集中以微博类短文本为主，所以验证集的平均长度和标签数都远大于测试集。

此外，本文还对测试集标签进行了深入分析和统计。结果显示，30.5%的文本不包含地名，而在文本中包含地名的前提下，大部分文本有且仅有1个主地域标签，占比89.8%，若这一数字取2，占比则达到98.0%。

集合	文本数	平均长度	平均标签数
主地域验证集	1226	1283.8	3.98
主地域测试集	1500	544.9	1.12

表 3. 主地域数据集的统计情况

参数名	值	参数名	值
word_dim	100	seg_dim	20
filter_width	3	lstm_dim	100
batch_size	120	optimizer	Adam
dropout	0.5	learing_rate	0.001

表 4. 实体识别模型的超参数设置

4.2 预处理和参数设置

本文采用了一些自然语言处理中通用的预处理方法，如对输入文本采取了ansj分词[5]和截断操作。为提升实体识别性能，采用了100维度的预训练词向量。为减少一些常见的地名干扰词，如“巴西龟”、“重庆鸡公煲”等对模型性能的影响，构建了停用词表，对文本进行去停用词操作。针对微博数据的特点，对url和表情符号等进行清洗。

对于命名实体识别模型IDLSTM+CRF，其初始化超参数如表4所示，本文使用Adam优化器，并初始化$\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-9}$。对于主地域抽取算法，本文设∂为1，β为2。

4.3 地名识别效果

我们采用精确率(Precision)、召回率(Recall)和F1-Score作为模型性能的衡量指标，其计算方法如下。

$$准确率(P) = \frac{正确识别的实体个数}{识别出的所有实体个数} * 100\%$$

$$召回率(R) = \frac{正确识别的实体个数}{样本中所有实体个数} * 100\%$$

$$F1 = \frac{2 * P * R}{P + R} * 100\%$$

IDLSTM+CRF的实体识别效果如表5所示，其中-attention代表不引入注意力机制，仅将两个特征抽取器输出的隐状态向量经过矩阵变换后拼接在一起，+attention代表引入注意力机制。两个基线模型效果相比，BiLSTM+CRF的效果要略好于IDCNN+CRF。

对比本文提出的模型，引入注意力机制的模型效果好于不引入注意力机制，且无论是否引入注意力机制，模型的效果均有所提升，其中表现最好的模型IDLSTM+CRF+attention，其F1-Score相较于BiLSTM+CRF在验证集上提升0.9%，在测试集上提升0.7%。若不引入注意力机制，IDLSTM+CRF模型在测试集上的精确度结果取得了最好成绩，但F1-Score在验证集上的效果相比于最好成绩下降了0.3%，在测试集上的效果也略有下降。这说明注意力机制可以动态的学习到分配权重给两个特征抽取器。上述实验结果验证了本文提出模型IDLSTM+CRF的有效性。

4.4 主地域抽取效果

我们在主地域验证集上和主地域测试集上进行了主地域抽取效果的测试。此外，为了进一步测试抽取的准确性，本文还线上随机抽取了500条文本进行人工联合评测，主地域抽取的效果如表6所示，可以看到本文提出的主地域抽取算法效果是比较好的，在表现最差的测试集上也能达到接近84.8%的抽取准确率。

我们对文本的来源进行了划分，并进行深入分析，如表7所示，在验证集上新闻文本的数量要远大于微博文本，在测试集上，两类文本的数量相仿，因为微博文本较短导致待识别的实体

[5] https://github.com/NLPchina/ansj_seg

命名实体识别验证集			
模型	精确率	召回率	F1-Score
BiLSTM+CRF	90.55%	88.48%	89.51%
IDCNN+CRF	89.79%	87.22%	88.49%
IDLSTM+CRF -attention	91.00%	89.23%	90.11%
IDLSTM+CRF +attention	**91.20%**	**89.64%**	**90.41%**
命名实体识别测试集			
模型	精确率	召回率	F1-Score
BiLSTM+CRF	89.08%	87.92%	88.49%
IDCNN+CRF	89.58%	87.69%	88.63%
IDLSTM+CRF -attention	**89.73%**	88.57%	89.15%
IDLSTM+CRF +attention	89.20%	**89.18%**	**89.19%**

表 5. 地名识别模型在命名实体识别验证集和测试集上的实验结果

集合	文本数量	准确率
主地域验证集	1226	91.7%
主地域测试集	1500	84.8%
人工评测数据集	426	87.3%

表 6. 主地域模型的抽取效果

集合	文本类型	数量	准确率
主地域验证集	新闻文本	938	91.4%
	微博文本	288	94.8%
主地域测试集	新闻文本	626	80.8%
	微博文本	874	89.6%

表 7. 主地域模型在不同文本上的抽取效果对比

数量较少，模型识别的难度要低于新闻文本，所以两个集合的微博文本识别效果都要好于新闻文本的识别效果，且测试集上微博文本的识别效果要高于新闻文本近9个百分点。

5 总结

本文首次提出了一套完整的主地域抽取系统，该系统主要由基于IDLSTM+Att+CRF的实体识别算法，基于统计信息的主地域抽取算法和基于最大匹配算法的主地域补全算法组成。经过真实场景下的测试，该系统能较好、较快的识别出非结构化文本中的主地域信息并进行自动补全，为舆情分析等下游任务打下基础。此外，为了解决当前主地域分析工作中缺少数量可观且统一的评测集合的痛点，我们以双工的形式标注并开源了两份主地域数据集，希望能推动该领域的进一步发展。

因采用流水线的方式构建我们的模型，每一个算法的性能都会影响下流任务的性能。为克服这一缺点，在未来的研究工作中，我们会探索联合式的主地域抽取算法。此外，受训练数据和方法所限，目前该系统只能对国内地名进行识别和补全，这在当今信息大爆炸的背景下是远远不够的，我们的实际测试时经常会碰到如“新山动物园”、“芽庄”等国外地名，因此未来考虑引入更加全面、精细的地名数据库和命名实体识别数据集，从而更全面的完成互联网文本所包含的主地域信息的抽取工作。

6 致谢

感谢匿名评审给出的意见与建议。本论文工作得到国家自然科学基金项目（项目批准号：6207611, U1908216, 61573294）的资助。

参考文献

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Maihemuti Maimaiti, Kahaerjiang Abiderexiti, Aishan Wumaier, Tuergen Yibulayin, WANG Lulu, et al. 2017. Crf与规则相结合的维吾尔文地名识别研究. 中文信息学报, 31(6):110–118.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent models of visual attention. *arXiv preprint arXiv:1406.6247*.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Emma Strubell, Patrick Verga, David Belanger, and Andrew McCallum. 2017. Fast and accurate entity recognition with iterated dilated convolutions. *arXiv preprint arXiv:1702.02098*.

俞鸿魁, 张华平, 刘群, 吕学强, and 施水才. 2006. 基于层叠隐马尔可夫模型的中文命名实体识别. 通信学报, 027(002):87–94.

刘佳琪and 罗永莲. 2019. 中文事件新闻的中国地名抽取算法研究. 信息与电脑, 000(015):53–54,57.

刘瑜, 张毅, 田原, and 薛露露. 2007. 广义地名及其本体研究. 地理地理信息科, 23(6):1–7.

唐旭日, 陈小荷, and 张雪英. 2010. 中文文本的地名解析方法研究. 武汉大学学报· 信息科学版, 35(8):930.

孙镇and 王惠临. 2010. 命名实体识别研究进展综述. 现代图书情报技术, 26(6).

张杰, 徐智婷, and 薛向阳. 2008. 融合多特征的最大熵汉语命名实体识别模型. 计算机研究与发展, 45(6).

张雪英, 闫国年, 李伯秋, and 陈文君. 2010. 基于规则的中文地址要素解析方法. 地球信息科学学报, 012(001):9–16.

智烈慧, 李仁杰, 傅学庆, and 郭风华. 2016. 众包旅游文本热度地名的共现挖掘. 测绘科学, 41(08):141–151.

曹春萍and 关鹏举. 2019. 基于e-cnn和blstm-crf的临床文本命名实体识别. 计算机应用研究, 036(012):3748–3751.

李妮, 关焕梅, 杨飘, and 董文永. 2020. 基于bert-idcnn-crf 的中文命名实体识别方法. 《山东大学学报(理学版)》, 55(1):102–109.

李照航, 郭风华, 李仁杰, 傅学庆, and 严正峰. 2015. 大量网络游记文本中热度地名提取方法与实证研究. 地理与地理信息科学, 31(1):68–73.

武惠, 吕立, and 于碧辉. 2019. 基于迁移学习和bilstm-crf 的中文命名实体识别. 小型微型计算机系统, 40(6):1142–1147.

石桢and 姚天. 2013. 一种基于统计和规则的核心地名抽取方法. 微型电脑应用, 29(02):56–59.

舒时立, 李锐, and 吴华意. 2019. 基于地名树的最佳空间尺度新闻事件地点提取方法. 武汉大学学报· 信息科学版, 44(9):1416–1422.

赵英, 占斌斌, 贾沛哲, and 李华英. 2017. 基于规则与词典的地址匹配算法. 北京测绘, 5:50–54.

钟翔, 高勇, and 邬伦. 2016. 基于链接分析的网页文本核心地名提取方法. 地球信息科学学报, 018(004):P.435–442.

马雷雷, 李宏伟, 梁汝鹏, 连世伟, and 龚竞. 2015. 基于地名本体的地名知识表达方法. 测绘科学技术学报, 32(03):305–309.

马雷雷, 李宏伟, 连世伟, 梁汝鹏, and 龚竞. 2016. 地名知识辅助的中文地名消歧方法. 地理与地理信息科学, 32(04):5–10.

黄炜, 黄建桥, and 李岳峰. 2019. 基于bilstm-crf的涉恐信息实体识别模型研究. 情报杂志, 38(12):149–156.