文章编号：1003-0077（2011）00-0000-00

旅游场景下的实体别名抽取联合模型*

杨一帆，陈文亮
（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘要：目前互联网中包含了大量的实体介绍文本，为实体知识构建提供了资源基础。别名作为实体的一种属性，是实体正式名称的不同表达，在知识图谱构建中具有重要意义。本文以景点介绍文本作为语料，结合不同别名描述方式提出别名标注策略，人工构建别名标注数据集。别名抽取可分为实体识别与关系分类两个子任务。本文提出基于深度学习的景点实体别名抽取联合模型，同时完成两个子任务。在本文构建的数据集上的实验结果表明，联合模型与流水线式处理模型相比性能有显著提高。

关键词：旅游景点；别名；联合模型；实体识别

中图分类号：TP391 **文献标识码：**A

Joint Model for entity Alias Extraction under Tourism Scene

YANG Yifan, CHEN Wenliang
(School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu, 215006, China)

Abstract : At present, the Internet contains a large amount of entity introduction texts, which provides a resource basis for the construction of entity knowledge. As an attribute of entity, an alias is a different expression of the official name of an entity and is of great significance in the construction of knowledge graphs. In this paper, the introduction text of the attraction is used as a corpus, and the alias annotation strategy is proposed in combination with different alias description methods to manually construct the alias annotation data set. Alias extraction can be divided into two subtasks: entity recognition and relation classification. This paper proposes a joint model of scenic entity alias extraction based on deep learning, and completes two subtasks simultaneously. The experimental results on the data set constructed in this paper show that the performance of the joint model are significantly improved compared with the pipelined model.

Key words: Tourist Attraction; Entity Alias; Joint model; Entity Recognition

* **收稿日期：**0000-00-00 **定稿日期：**0000-00-00
基金项目：国家自然科学基金（61572338，61876115）
作者简介：杨一帆（1995），男，硕士研究生，主要研究方向自然语言处理；陈文亮（1977），男，教授，主要研究方向自然语言处理。

1 引言

命名实体识别[1]与关系分类[2]均为自然语言处理中较为经典的任务，而关系抽取[3]则需要两者的结合。实体别名抽取与关系抽取类似，该任务不仅需要从自由文本中获取实体对，还需判断相应的实体对是否存在别名关系。

近年来，随着生活水平提高以及信息化水平发展，智能出行也成为了人们生活的一部分。然而对于旅游景点名称查询常常伴随着差异，例如用户希望查询景点“鸟巢”的相关信息，实际查询目标为“国家体育场”，该景点被称为“鸟巢”是因其外形类似鸟的巢穴；而“故宫”与“紫禁城”，两者存在别名关系，但若无任何背景知识，单纯依靠字面理解，无法将两者联系在一起。因此，从背景知识中挖掘出景点别名，构建旅游领域别名库，不仅能够方便用户查询，提供丰富准确的信息，亦能够用于知识图谱属性补全[4]，为文本处理提供可靠的结构化知识。

图1 候选别名实体获取示例

在本任务中，我们通过爬虫获取了大量自由文本，其中每一文本均为某一景点的描述文本。我们希望从对应的描述文本中，获取目标主体的别名。根据别名的分布规律，我们将其分为两种类型，并在标注过程中有所区分。本文别名识别共分为两步：文本中的景点实体识别，以及识别结果与当前主体是否具有别名关系的判断分类。

如图1所示，从描述文本中，我们能够获取大量景点实体和昵称作为主体的别名候选集合。该文本中，在未得知主体的情况下，“新都桥”、“‘摄影家走廊’”、“偶曲河”、“‘摄影家天堂’”等实体均有可能与主体指向同一景点。当得知该描述文本的主体为“白玉偶曲河”时，上述实体中仅“偶曲河”和“‘摄影家天堂’”符合别名条件。最终主体“白玉偶曲河”的别名集合为{偶曲河，“摄影家天堂”}。

然而一些主体的描述文本不包含其本身，为解决此类描述文本主体缺失的问题，本文提出一种针对该问题的别名识别模型。模型以拼接卷积的方式，将主体信息添加至文本中进行编码，加强其在文本中的主导作用，由此帮助确定目标主体的别名并过滤无关实体。本文的贡献点在于：(1)提出一种基于当前任务的别名标注策略，并构建出一份景点别名的人工标注语料，可用于将来研究；(2)针对别名抽取任务的特点，提出新的信息编码方法，构建了与标注策略相结合的联合模型；(3)将语言模型ELMo[5]应用于模型中，进一步提升别名抽取系统的性能。

2 相关工作

目前鲜有专门针对别名属性的抽取方案，与本文最相关的任务即关系抽取。针对关系抽取任务一般有两种建模框架：流水线式处理模型(Pipeline Method)以及联合模型(Joint Model)。流水线式处理模型依序分别处理两个子任务，即先对文本进行命名实体识别(NER)，再对识别实体进一步关系分类(RC)，在本文场景下，关系分类模型则用于判断识别实体与目标主体是否具有别名关系。联合模型则将两个子任务合并同时处理，减少分步处理带来的错误。

命名实体识别(Named Entity Recognition, NER)作为信息抽取的重要基础，具有广泛的应用，能够从文本中识别人名、地名等类型实体，目前一般将其作为序列标注问题处理。传统用于NER的线性统计模型有条件随机场(CRF)[6]和隐马尔科夫模型(HMM)[7]等，该类模型通常需要人工构建复杂的特征模板。随着深度学习的发展，循环神经网络(RNN)的加入提升了该任务的效果，RNN较其它神经网络模型效果显著，并且不断变形发展。Collobert等[8]最早通过深度神经网络进行命名实体识别。Lample等[9]提出了基于BiLSTM与CRF解码层得到序列标签，在该任务上取得了较好的效果。Ma等[10]提出BiLSTM-CNNs-CRF的端到端识别模型，能够很好地适用于各种序列标注任务。

关系分类(Relation Classification, RC)通常是一个多分类任务。现有的分类模型可分为基于手工特征的方法和基于神经网络的方法。基于手工特征的传统方法采用一系列NLP工具完成特征提取，十分依赖特征设计的质量。Kambhatla[11]等基于词汇、句法及语义特征，并结合最大熵模型实现分类。目前常见用于分类的深度学习模型有卷积神经网络CNN及循环神经网络RNN等。Zeng

等人[12]提出通过CNN获取词汇级别及句子级别的特征，以此进行分类。Xu等[13]人提出基于LSTM的最短依赖路径作为特征实现关系分类。Cai等人[14]将双向RNN与CNN相结合，提出双向循环卷积神经网络(BRCNN)，同样基于最短依赖路径，在关系分类任务上有所突破。

近年来，针对实体关系抽取的联合模型也有了一定的发展。Zheng等[15]提出通过构建底层参数共享的混合神经网络，实现实体识别及关系分类两个子任务之间的依赖。Miwa等[16]采用基于序列层和依赖树的结构信息对文本进行关系抽取。Zheng等[17]提出一种全新的标注策略，将原有的实体识别任务和关系分类任务完全变成一个序列标注任务，并通过端到端的神经网络模型直接获取实体关系三元组。

而上述方法均以关系实体同时存在于文本中为前提[18]，实际本文数据的描述文本存在较多主体缺失的情况，这给关系分类带来困难。针对该问题，本文提出将主体特征信息嵌入句中进行编码，明确别名的目标，并以联合方式建模，实现针对主体的别名抽取。

表1 常见的别名类别

别名类别	主体	示例	实体类型
指向性昵称	故宫	故宫又叫做【紫禁城】…	AE
新旧昵称	中国国家馆	…后来更名为【中华艺术宫】	AE
	上海动物园	…原名为【西郊公园】	AE
誉名	苏州	…有【“人间天堂”】的美誉	AE
实体缩写	苏州大学	【苏大】校园内风景优美…	SE
添加或省略位置	国家图书馆	【中国国家图书馆】是中国…	SE
字符重叠	王府井天主堂	【王府井教堂】位于…	SE

3 数据

本文以爬取的旅游景点描述文本作为语料进行人工标注，共计5,000条记录[①]。每条数据记录分为两部分：目标主体和该主体的描述文本，格式如图1所示。其中大约有1/3的数据记录，它们的主体并没有出现在其描述文本中，这给别名识别带来一定的困难。

3.1 别名定义

别名，是指正式名称以外的其他名称，它们均指向同一种事物。在本文中，我们规定以数据中的主体作为正式名称，对于那些不同于主体但与主体表示同一事物的实体名称，均作为该主体的别名。由于语言的多样性，别名也拥有多种形式，常见的别名类别有实体缩写、昵称等。表1列出几种归纳后的别名类别。

3.2 数据处理

常规字符的切分一般将每个字符作为单个token[19]，或是通过预分词[20]，将每一切分结果作为单个token。由于描述文本以介绍为主，既包含大量中文字符，也拥有不少英文字符，以上两种token切分方式均会带来一些问题。

如图2所示，例1文本中英文单词较多，若单个字符作为token，则字母将占用大量序列长度，这是毫无意义的。而通过预分词，原有实体可能被切分并与附近其他字符另组词，如原实体为“虹

例1 美国圣马力诺亨廷顿图书馆(The Huntington Library, Art Collections and Botanical Gardens)是……

例2 虹之松/原位于/唐津市/的/唐津湾/沿岸

图2 两种token切分方式问题示例

美/国/圣/马/力/诺/亨/廷/顿/图/书/馆/
(/The/Huntington/Library/,/Art/Collections/
and/Botanical/ /Gardens/)/是/……

图3 本文token切分方式示例

之松原”，切分后“原”与“位于”组成了新的词组，并且由分词引发的该类错误将不可逆。

基于此，本文结合两者：对于中文字符，以单字符作为token；对于英文单词，以正则匹配的方式，将获取的单词作为单个token。示例见图3。

3.3 几点假设

通过对描述文本数据的观察研究，我们认为候选别名实体基本满足以下条件之一：1)该别名实体为景点实体且在句中为主语。2)该别名实体具有上下文指代信息。根据上述条件，我们将能够作为别名的实体分为两类，以SE(Subject Entity)表示句中景点的主语实体，AE(Alias Entity)表示具有上下文指代的昵称实体，具体如表1所示。

[①]本文实验数据链接：
https://github.com/LimKim/Tourist-Attraction-Alias

针对本文场景的别名抽取任务，我们令主体为e，描述文本为d，描述文本d中所有主语实体(SE)集合为S_d，具有上下文指向的所有昵称实体(AE)集合为A_d。基于此，我们提出以下几点假设，明确别名抽取的范围，并对模型识别结果进行后处理，在此基础上实现别名抽取。

假设1　若e与d相对应，则S_d中至少存在一个主语实体s，使得$s=e$或s与e有较高的字符重叠。

由于d为主体e的描述文本，一般情况下d应当围绕e展开介绍，因此必存在以e或与e字符相似的实体作为主语的语句。若不存在满足条件的主语实体s，则无法认为d是主体e的描述文本，这与“e与d相对应”相悖。

若“苏州大学”的描述文本中未出现过其主体或类似“苏大”的相似实体，则认为该段描述文本与主体“苏州大学”关联不大，无法从中获取有关主体的有用信息。

假设2　若e与d相对应，存在集合$\widetilde{S_d}\subseteq S_d$，且$\widetilde{S_d}$中每一实体$s$均满足与$e$具有较高字符重叠，则对于任一$s\in\widetilde{S_d}$，均有可能为$e$的别名。

若存在满足与主体e具有较高字符重叠的主语实体集合$\widetilde{S_d}$，则该集合内任一实体s一定与主体e相关，因此可以令该集合作为主体e的别名候选集合。然而并不是仅满足字符重叠就能够认为s是e的别名。

例如主体“乾坤山”，其描述文本中识别出实体“乾坤寺”，该类字符重叠情况并非我们所需要的，两者只是相关实体，不存在别名关系。因此，我们希望通过后续训练出合适的分类器模型，自动判别该类问题。

假设3　若e与d相对应，存在a属于A_d，且a的关系主语为e，则a是e的别名。

由于拥有强烈的指向性文本，具有较高的可信度，因此认为a为e的别名。

如描述文本“故宫又叫做紫禁城…”，该类情况认为“紫禁城”即主体“故宫”的别名。

假设4　若e与d相对应，存在实体m，满足$m\notin S_d$且$m\notin A_d$，则认为m不是主体e的别名。

在没有任何外部知识背景下，我们无法认为某个与主体e既无字符重叠且无上下文指向的实体m是主体e的别名。

若文本中同时出现“故宫”与“紫禁城”，但文中未提及“故宫又叫做紫禁城”等近义句，而是分别陈述两者，则对于读者而言，依据仅有知识无法得知两者关系，因此不认为“紫禁城”与“故宫”存在别名关系。

3.4 标注策略

对于常见的序列标注任务，一般使用IOB2标注体系[21]，同时也会加入类型后缀，如人名(PER)、地名(LOC)等。其中B代表被标注实体的初始词，I代表除初始词外其他部分的词，O代表未被标注的词。由于当前场景较为特殊，文本中主体缺失的情况颇多，因此只标记出文本中主体的别名实体。

针对本文场景下的别名抽取任务，常规标注仅标记出满足别名条件的正确实体，标注结果如表2所示。而本文在标注时将根据对候选别名实体的分析，标记出实体的类型。同时该方案也会标记出错误的别名实体，即非目标主体的别名。最终我们按照该标注策略将满足条件的候选实体均标记出来，结果如表3所示。

表2 “白玉偶曲河”部分描述文本常规标注结果

字词	…	那	么	偶	曲	河	就	是	“	摄	影	家	天	堂	”
标签	O	O	O	O	O	O	O	O	B	I	I	I	I	I	I

我们安排了2位标注员并行对该数据进行标注，平均每100条数据标注时间约为1小时，总标注量约为50小时-人，最终获得5,000条人工标注数据记录。其中描述文本主体缺失的记录数量为1,789条，描述文本平均长度约为43.85个字符，平均每条数据记录中标记类型为SE的实体个数约为1.44个，类型为AE的实体个数约为0.25个。

表3 “白玉偶曲河”部分描述文本本文标注结果

字词	标签	字词	标签
如	O	,	O
果	O	那	O
说	O	么	O
新	B-SE_False	偶	B-SE_True
都	I-SE_False	曲	I-SE_True
桥	I-SE_False	河	I-SE_True
被	O	就	O
称	O	是	O
为	O	“	B-AE_True
“	B-AE_False	摄	I-AE_True
摄	I-AE_False	影	I-AE_True
影	I-AE_False	家	I-AE_True
家	I-AE_False	天	I-AE_True
走	I-AE_False	堂	I-AE_True
廊	I-AE_False	”	I-AE_True
”	I-AE_False	。	O

4 模型框架

在别名抽取任务中，不仅要识别出文本中可

能与目标主体具有别名关系的候选实体，而且需要进一步判断是否为目标主体的别名。本文采用命名实体识别与别名判断分类器联合学习策略，即同时进行候选实体识别和别名关系判断。在别名关系判断时，为解决文本中目标主体缺失的问题，我们将目标主体信息添加到句子中进行编码。

图4是联合模型框架示意图，本文模型分为两部分。第一部分是候选实体识别，第二部分是别名关系判断。实体识别时，我们获取每个token的B/I/O标签，以及类型标签SE/AE。随后，我们针对识别出来的候选实体，分类判断其是否为当前主体的别名。对于分类标签为True的实体，我们认为该实体与主体存在别名关系；同样的，对于分类标签为False的实体，认为该实体并非主体的别名。

4.1 实体识别

模型第一层为数据表示层。对于一段输入序列，该层将每一字符映射为向量表示。本文采用了近年来较为有效的ELMo模型获取句子级别的向量输入表示。ELMo(Embedding from Language Models)是一种新型深度语境化词特征，能够针对词在语境中的变化建模，产生丰富的词语表征。

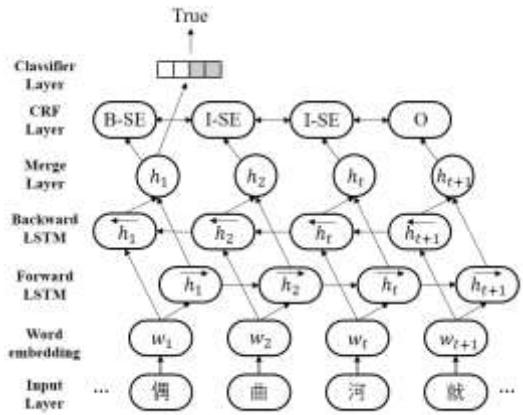

图4 联合模型框架示意

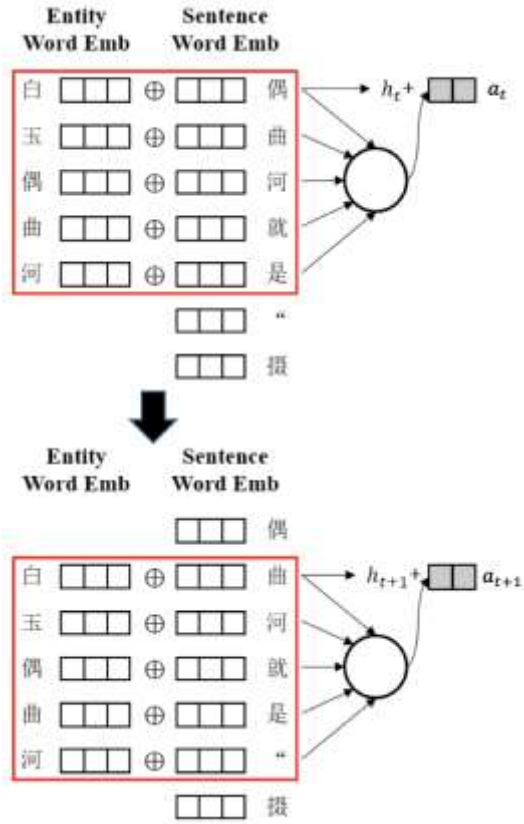

图5 主体嵌入操作示意

其次是BiLSTM层。LSTM在RNN的基础上增加了输入门、遗忘门、输出门，能够根据信息特征的重要性进行保留与遗忘，解决了神经网络中长序列依赖问题，主要公式如下所示：

$$i_t = \sigma(w_t \cdot W_{xi} + h_{t-1} \cdot W_{hi} + b_i) \quad (1)$$

$$f_t = \sigma(w_t \cdot W_{xf} + h_{t-1} \cdot W_{hf} + b_f) \quad (2)$$

$$o_t = \sigma(w_t \cdot W_{xo} + h_{t-1} \cdot W_{ho} + b_o) \quad (3)$$

$$\widetilde{c_t} = \tanh(w_t \cdot W_{xc} + h_{t-1} \cdot W_{hc} + b_c) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \widetilde{c_t} \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

其中，w_t代表t时刻的输入，σ代表$sigmoid$函数，$\odot$代表元素点积运算，i_t，f_t，o_t分别为t时刻的输入门、输出门及遗忘门，$\widetilde{c_t}$，c_t，h_t分别为t时刻的候选细胞状态，细胞状态及隐层状态。对于每一单位输入w_t，LSTM层会考虑从w_1至w_t的上下文信息，令该信息输出为$\overrightarrow{h_t}$；同样的，对于另一向的隐层信息输出表示为$\overleftarrow{h_t}$。两者进行拼接后，最终的隐层输出为$h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$。

CRF层能够很好地合理化标签序列。对于一组输入序列$X = (x_1, x_2, \dots, x_n)$，其对应的预测标签序列$Y = (y_1, y_2, \dots, y_n)$，定义它的得分如下所示：

$$Score(X, Y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \quad (7)$$

其中，A为转移得分矩阵，$A_{i,j}$表示从标签i转

移至标签j的得分，$P_{i,j}$表示第i个输入状态被标注为第j个标签的得分。y_0和y_n分别为标签序列的起始标签和结束标签，假设标签个数为k，则转移得分矩阵A为$k+2$阶方阵。经过归一化，得到每个标签序列的条件概率，表示如下：

$$p(y|X)=\frac{e^{Score(X,y)}}{\sum_{\tilde{y}\in Y_X}e^{Score(X,\tilde{y})}} \tag{8}$$

$$L_1=max\sum_{d=1}^{|\mathbb{D}|}p(y_d|X_d)$$

其中$|\mathbb{D}|$为训练集的数据总个数，Y_X表示输入序列X_d所有可能的标签序列集合，y_d为真实的标签序列。训练目标即最大化上式真实标签序列的对数似然概率，目标函数即为L_1。最终对自由文本标注时，取满足下式的结果y^*作为最佳预测标签序列。

$$y^*=\arg\max_{y\in Y_X}Score(x,\tilde{y}) \tag{9}$$

4.2 别名判断分类

至此，模型获取了所有的候选实体。为了加深主体在句中主导性，增强模型对实体间字符重叠的特征感知，我们将主体从句首词开始滑动拼接，对拼接后的结果进行特征提取。随后，我们将特征提取结果与当前字词的隐层输出相拼接，共同预测当前部分实体的分类标签，即判断当前实体是否与主体具有别名关系。

如图5所示，对于主体输入序列$W^+=(w_1^+,w_2^+,\dots,w_m^+)$及描述文本输入序列$W=(w_1,w_2,\dots,w_n)$，有如下操作：

$$a_t=ReLU\left(\sum_{i=1}^{m}W_i\cdot(w_i^+\oplus w_{t+i-1})\right) \tag{10}$$

其中$\oplus$为向量拼接运算，$ReLU$为激活函数，m,n分别为主体长度和描述文本序列长度，且$m\ll n$。则a_t表示以w_t为首的候选实体与目标主体计算后的特征结果，我们将其与当前时刻的隐层输出h_t拼接后，进行分类预测结果：

$$Output_t=h_t\oplus a_t \tag{11}$$

$$y_t=W_y\cdot Output_t+b_y \tag{12}$$

$$p_t^f=\frac{exp(y_t^f)}{\sum_{\tilde{f}\in F}exp\left(y_t^{\tilde{f}}\right)} \tag{13}$$

其中W_y为softmax矩阵，p_t^f表示t时刻分类标签为f的概率，$F=\{True,False\}$。此时$Output$中既包含与目标主体的相似度特征，同时注意到目标主体的上下文特征，因此以$Output$作为分类的特征向量。

对于一段描述文本，模型能够从中识别出多个实体作为别名候选。因此我们需要对每一识别实体，即不同位置的部分描述文本进行分类。为方便模型并行处理，我们利用每一实体的首字分类标签作为当前位置识别实体的分类结果，并以序列标注的方式完成所有实体的分类。忽略实体首字以外的标签带来的影响，我们将目标函数定义为如下所示：

$$L_2=max\sum_{d=1}^{|\mathbb{D}|}\sum_{t=1}^{n}log\left(p_t^{real}|X_d\right)\cdot r(B),$$

$$r(B)=\begin{cases}1, & if\ real\ tag='B'\\ 0, & if\ real\ tag\neq'B'\end{cases}$$

其中p_t^{real}为t时刻真实标签对应的概率值，$|\mathbb{D}|$为训练集的数据总个数，$r(B)$用于忽略实体首字以外的标签带来的影响。当实体首字的分类标签确定后，以该结果作为当前实体分类结果。

由于共享同一编码层Merge Layer，两个任务在训练时能够互相影响，因此最终的联合目标函数即为两者目标函数合并结果。我们依据两者的目标函数获得各自的损失函数，并以两者的加权和作为联合损失函数，实现多任务联合训练。

$$L=L_1+L_2$$

5 实验

本节主要介绍模型的参数设置、实验评估方式、实验结果以及错误实例分析等。

5.1 实验设置

本文实验采用准确率(Precision)、召回率(Recall)及F1值评估实验结果。以人工标注的5,000条记录作为实验数据，平均切分为10份并由1至10编号，固定以1号数据集作为测试集，在2号到10号数据集中轮流取其中1份作为验证集，并将剩余8份用作训练集，以此进行3次实验。最后依次采用验证集最佳效果模型，获取测试集的实验结果，并取3次实验平均结果作为最终实验结果。由于本文任务是获取主体的别名，因此实验评估时只考虑与主体存在字符差异的识别结果，对于与主体一致的识别结果，不将其纳入评估范围内。

本文采用BiLSTM-CRF进行实体识别，随后利用实体首字符特征向量进行分类，由此获取实体最终标签。模型的参数设置如表4所示。

表4 模型超参数设置

参数	值
ELMo字向量维度	600
隐层维度	256
优化函数	RMSprop
学习率	0.001
batch_size	8
dropout	0.5

5.2 实验分析

本文实验结果分为两部分，第一部分为流水线式处理模型的实验结果，第二部分为联合模型方法的实验结果。

基准实验采用条件随机场进行实体识别并以最大熵模型分类(CRF+ME)[22]，该方案首先识别出所有的主语实体和昵称实体作为目标主体的别名候选，再针对是否与目标主体具有别名关系进行分类。其次在该基础上改进，以BiLSTM-CRF[9]的方式进行实体识别，分类器则采用CNN模型。

联合模型LSTM-LSTM-Bias采用Zheng等[17]人提出的方法。该方法通过BiLSTM对输入进行编码再以LSTM解码，并根据标签类型给定不同的权重，使得真实标签为非O的token权重更大，令模型更加关注实体所处位置的特征信息。

联合模型BiLSTM-CRF+C&C采用了常规的标注规范，通过本文方法共享同一编码层将两个任务合并处理。在实体识别的同时加入主体拼接卷积后的特征，以此进行分类过滤非目标主体的别名(Concat&Classify, C&C)。

LS(Labeling strategy)即本文的标注策略，AP(Assumptions)即本文基于场景提出的假设约束。在上述的别名识别联合模型中，以本文的LS替代常规的标注规范，使得序列标注更好地注意不同类型的候选实体，减少识别步骤产生的误差。本文的最终模型通过LS&AP相结合，并基于合理假设对模型最终结果进行纠正与后处理。在此基础上，本文以基于当前语料训练的ELMo字向量代替原有的随机字向量。

5.3 实验结果

具体实验结果如表5所示。在流水线式处理模型中，传统模型结果一般。而以深度学习模型代替传统模型，在原有基础上实验效果有所提高。经果综合对比，我们认为流水线式处理模型整体效果与联合模型相比有所不足。

联合模型LSTM-LSTM-Bias中，该模型利用改进Cell后的单层LSTM代替了传统的CRF层，使其在识别时出现一些非法标签序列，导致该模型与其他联合模型相比，准确率和召回率均有不足。

表5 不同模型实验结果对比

Methods	Precision/%	Recall/%	F1/%
CRF+ME	66.95	46.85	55.12
BiLSTM-CRF+CNN	65.67	59.16	62.24
LSTM-LSTM-Bias	64.61	61.42	62.97
BiLSTM-CRF+C&C	69.14	**66.67**	67.88
LS&AP+BiLSTM-CRF+C&C	75.60	63.01	68.73
LS&AP+BiLSTM-CRF+C&C+ELMo	**75.90**	64.98	**70.02**

而对于联合模型BiLSTM-CRF+C&C，该模型能够并行处理两个子任务，识别出大量候选实体，并在分类时也有着较高的准确率。但通过评估，仍存在一些错误识别结果，与本文的假设相悖。

在BiLSTM-CRF+C&C基础上，模型结合了LS&AP，即本文的标注策略以及提出的假设。可以看出，该方案有效地过滤了部分被模型误识别为正确的别名。虽然该方案同时过滤了少量正确别名，部分召回率有所牺牲，但准确率与之前的联合模型相比均有所提高，且整体效果优于前者。在当前方案基础上，我们使用基于当前语料训练出的ELMo字向量代替当前字向量，最终实验结果达到了当前任务的最优结果。

实验结果表明，针对别名抽取任务而言，联合模型的效果优于流水线式处理模型，而主体特征信息也有助于我们过滤掉错误的识别结果。

5.4 错误分析

针对当前任务，本文模型的实验效果相对有所提高。然而在此基础上实验仍有一定提升空间，以下是针对当前模型的一些错误分析：

上下文表述形式特殊。实验语料由爬虫所获得，一些描述文本存在特殊句式，如文言文等。例如：“佩枯错一作佩估错……”，“一作”可被翻译为“也叫作”，但由于该表达方式出现频率较低，上下文语义特征不明显，导致实体无法被识别。

语言种类的差异性。本文语料中不乏国外旅游景点介绍，因此文本中常伴有其他语言。例如：描述文本“Phuket Zoo是普吉岛上的一个小型动物园……”的主体是“普吉岛动物园”，而描述文本一直以“Phuket Zoo”表示，根据本文方法无法获取两者的相似特征与上下文特征。这对识别“Phuket Zoo”产生负面影响，分类时因无字符重叠可能会导致判断出现偏差。

实体命名生僻。描述文本“巴松错在藏语里意为‘绿色的水’，是红教的著名神湖……”中，“巴松错”即“巴松错湖”，但由于该实体并非以常见的景点名称字符命名，且上下文界限模糊，因此该实体并未被当作景点实体识别出来。

实体枚举过多。一些描述文本会枚举大量实体，导致位于文本尾部实体的上下文特征弱化，无法被成功识别。如描述文本“小仓城，别称胜

山城、胜野城、指月城、涌金城和鲤之城……”，其中类型为AE的实体有5个，但在实际识别过程中无法全部识别，时常发生实体识别缺失，或将多个实体共同识别为一个整体等情况。

6 结束语

本文以爬取的旅游景点描述文本作为语料，根据别名出现的规律设计人工标注规范，提出合理假设加以约束并在此基础上进行人工标注，最终获得了景点别名标注语料。针对本文任务的特点，本文模型将目标主体特征以拼接卷积的方式嵌入到句子表达中，实现针对主体的别名抽取，取得了较好的效果。

旅游景点通常拥有较多别名。获取景点别名不仅能够方便人们更好地查询该景点相关信息，亦能提高知识库的覆盖度，为后续的自然语言处理提供良好的知识背景。本文提出的联合模型与流水线式处理模型相比有一定提高，但仍有较大提高空间。未来将考虑加入语义信息特征或者其他语言模型（如BERT），进一步提高识别的效果。

参考文献

[1] Chinchor N, Robinson P. MUC-7 named entity task definition [C]//Proceedings of the 7th Conference on Message Understanding. 1997, 29: 1-21.

[2] Hendrickx I, Kim S N, Kozareva Z, et al. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals[C] //Proceedings of the Workshop on Semantic Evalua-tions: Recent Achieve-ments and Future Directions. Association for Computa-tional Linguistics, 2009: 94-99.

[3] Bach N, Badaskar S. A review of relation extraction[J]. Literature review for Language and Statistics II, 2007, 2.

[4] Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge[C]// Proceedings of the 16th interna-tional conference on World Wide Web. ACM, 2007: 697-706.

[5] Peters M E, Neumann M, Iyyer M, et al. Deep context-ualized word representations[J]. arXiv preprint arXiv: 1802.05365, 2018.

[6] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[J]. 2001.

[7] Passos A, Kumar V, McCallum A. Lexicon infused phrase embeddings for named entity resolution[J]. arXiv preprint arXiv:1404.5367, 2014:78-86

[8] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th international conference on Machine learning. ACM, 2008: 160-167.

[9] G. Lample, M. Ballesteros, S. Subramanian, K. Kawaka-mi, C. Dyer, Neural architecttures for named entity recognition: Proceedings of the Annual Conference of the North American Chapter of the Association for Com-putational Linguistics, 2016.

[10] Ma X, Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf[J]. arXiv preprint arXiv: 1603.01354, 2016.

[11] Nanda Kambhatla. 2004. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. Proceedings of the 43th ACL international conference. page 22.

[12] D. J. Zeng, K. Liu, S. W. Lai, G. Y. Zhou, and J. Zhao, Relation classification via convolutional deep neural network, in Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Paper, Dublin, Ireland, 2014, pp. 2335–2344.

[13] Xu Y, Mou L, Li G, et al. Classifying relations via long short term memory networks along shortest dependency paths[C]//Proceedings of the 2015 conference on empirical methods in natural language processing. 2015: 1785-1794.

[14] Cai R, Zhang X, Wang H. Bidirectional recurrent convolutional neural network for relation classifica-tion[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016, 1: 756-765.

[15] Zheng S, Hao Y, Lu D, et al. Joint entity and relation extraction based on a hybrid neural network[J]. Neurocomputing, 2017, 257: 59-66.

[16] Miwa M, Bansal M. End-to-end relation extraction using lstms on sequences and tree structures [J].arXiv preprint arXiv:1601.00770, 2016.

[17] Zheng S, Wang F, Bao H, et al. Joint extraction of entities and relations based on a novel tagging scheme[J]. arXiv preprint arXiv:1706.05075, 2017.

[18] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data[C]//Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. Association for Computational Linguistics, 2009: 1003-1011.

[19] 奉国和,郑伟.国内中文自动分词技术研究综述[J].图书情报工作,2011,55(02):41-45.

[20] Sang E F, Veenstra J. Representing text chunks[C] //Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 1999: 173-179.

[21] Harrington P. Machine learning in action[M]. Greenwich: Manning, 2012.

[22] Li Q, Ji H. Incremental joint extraction of entity mentions and relations[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014, 1: 402-412.