面向先秦典籍的历史事件基本实体构件自动识别研究*

王东波　高瑞卿　沈　思　李　斌

摘　要　结合数字人文的数据获取、标注和分析方法,识别和挖掘先秦典籍中历史事件基本实体构件具有重要的推广和使用价值。本文将先秦时期极具代表性的《公羊传》《谷梁传》《左氏春秋》《吕氏春秋》《晏子春秋》等作为处理语料,对其中的人名、地名、时间实体等进行内部数量统计和外部特征分析,构建特征模板。在已有的465,197个词汇的基础上进行实体抽取训练与测试,选出人名、地名、时间实体识别效果的调和平均数最大(87.37%)的模型,并将其应用于《国语》语料以便检验识别效果,同时将以上过程进行可视化展现。图8。表11。参考文献13。

关键词　条件随机场　数字人文　命名实体　先秦语料库

Research on Automatic Recognition of Basic Entity Component of Historic Events for Pre-Qin Classics

Wang Dongbo　Gao Ruiqing　Shen Si　Li Bin

Abstract: Combined with the method of data acquisition, annotation and analysis of digital humanities, it is of great promotion and application value to understand the ancient Chinese language and excavate the basic physical components of historical events in the pre-Qin classics. Using such representative of the pre-Qin period collections as *Gongyangzhuan*, *Guliangzhuan*, *Zuo's Spring and Autumn*, *Lv's Spring and Autumn* and *Yanzi's Spring and Autumn*, this article analyzes the person entities, site entities and time entities' inside and external feature statistics to build a feature template. On the scale of 465,197 words, we extract the entities of the training and testing, the best average number of recognition effect of selecting the person entities, site entities, time entities reached to 87.37%, and apply it in the *Guoyu* corpus to verify the recognition effect, finally we visualize the above process. 8 figs. 11 tabs. 13 refs.

Keywords: Conditional Random Fields; Digital Humanities; Named Entity; Pre-Qin Corpus

1　引言

近几年,大规模的古籍整理项目普遍缺乏“互联网思维”,几乎没有考虑借鉴数字人文的思路和方法,没有充分利用信息技术的巨大优势[1]。国内对古籍文本的开发与利用基本还是采取传统的研究方法和模式,规模庞大的古汉语典籍数据与较低的深度利用度之间的矛盾比较突出[2]。随着数字人文的迅速发展[3,4],业界对古汉语典籍的数字化、深度挖掘的需求越来越迫切。事件抽取已经成为现代信息挖掘研究领域的一个新的重要课题,但目前对古汉语典籍中事件命名实体自动识别的研究还相对较少。同时,先秦时期的典籍体现了华夏范围内的基本时代特征,其内容反映了春秋时期的各种社会关系和文化现象。选取历史或具有历史性质的先秦典籍进行实体识别的探究正是在上述背景下的一种尝试。本文所谓的历史事件基本实体构件是指构成一个事件的人物、地点和时间这三种实体。

* 本文系国家社会科学基金重大项目“基于《汉学引得丛刊》的典籍知识库构建及人文计算研究”(项目编号:15ZDB127)和国家自然科学基金面上项目“基于典籍引得的句法级汉英平行语料库构建及人文计算研究”(项目编号:71673143)的研究成果之一。

实体识别主要是从文本中抽取出我们称为“实体”的特定事实信息,如时间实体、地名实体、动词等等[5]。目前,利用条件随机场进行实体识别或抽取的研究更多集中于现代汉语的文本上,比较有代表性的研究如:徐元子等人针对新闻事件中展开的网络评论,提出一种基于条件随机场的网络评论与新闻事件中命名实体的匹配方法[6];陈箫箫等运用条件随机场模型完成事件抽取中的序列标记任务,非监督分类方法选用了LDA主题模型,实现一个事件抽取和分类系统,并且命名实体识别和事件短语抽取均取得较高的准确率和召回率[7];陈慧炜在分析案件文本特点的基础上,采用知识表辅助机器学习为主要的实验方法,并结合利用了条件随机场(CRF)统计模型,以刑事类案件文本为对象进行了信息抽取研究[8];景悦诚实现了“抓取微博—去除噪音—句子分割(分词,词性标注,命名实体识别,依存句法关系)—人工标注—机器学习—事件发掘”的完整流程,对新浪微博进行文本挖掘[9]。在古代汉语文本上进行实体识别和事件抽取的研究相对较少。李劲等人基于条件随机场提出一个概率模型,用于从大量个人简历中抽取有用信息,并进一步挖掘个人简历中所包含事件之间的关联关系[10];黄水清等基于人工标注的《左传》语料,将条件随机场模型与最大熵模型在地名实体抽取效果上相比较,验证条件随机场模型对地名实体的识别效果优于最大熵模型[11]。上述基于条件随机场的实体抽取研究已取得令人满意的效果,为本文利用条件随机场进行历史事件基本实体构件自动识别的研究奠定了基础。

本文以文本挖掘中的条件随机场模型为基础,对由《公羊传》《谷梁传》《左氏春秋》《吕氏春秋》《晏子春秋》等文本组成的先秦语料库中能够构成历史事件基本实体构件的人名、地名、时间实体进行抽取实验,为探究面向先秦典籍的历史事件提供基本的实体数据素材。

2 语料库及模型简介

在对《公羊传》《谷梁传》《左氏春秋》《吕氏春秋》《晏子春秋》五部先秦典籍进行人工自动分词、词性和实体标注的基础上①,本文构建了具有历史性质的先秦典籍语料库,总词汇规模达到465,197个。语料中需要被识别的人名、地名、时间实体分别被标注成“nr”“ns”“t”,语料库的具体样例如表1所示。

表1 典籍文本标注前后对照样例

原始文本	标注后的文本
段失子弟之道矣,	段/nr 失/v 子/n 弟/n 之/u 道/n 矣/y ,/w
賤段而甚鄭伯也。	賤/vy 段/nr 而/c 甚/v 鄭伯/nr 也/y 。/w
何甚乎鄭伯?	何/r 甚/v 乎/v 鄭伯/nr ?/w
甚鄭伯之處心積慮成於殺也。	甚/v 鄭伯/nr 之/u 處/v 心/n 積/v 慮/n 成/v 於/p 殺/v 也/y 。/w
于鄢、遠也,	于/p 鄢/ns 、/w 遠/a 也/y ,/w
猶曰取之其母之懷中而殺之云爾,	猶/d 曰/v 取/v 之/r 其/r 母/n 之/u 懷/n 中/f 而/c 殺/v 之/r 云/y 爾/r ,/w
甚之也。	甚/a 之/r 也/y 。/w
然則為鄭伯者宜奈何?	然/c 則/c 為/v 鄭伯/nr 者/r 宜/d 奈/v 何/r ?/w

① 本文所有关于先秦典籍的自动分词、词性和实体标注均由南京师范大学文学院古典文献与计算语言学专业的学者和博士研究生经过三轮标注与校验而完成。

续表

原始文本	标注后的文本
緩追逸賊，	緩/za　追/v　逸/v　賊/n　，/w
親親之道也。	親/v　親/n　之/u　道/n　也/y　。/w
秋七月，	秋/n　七月/t　，/w

本文所有的实验均是基于标注后的共有465,179个词汇的先秦典籍语料库进行的,主要是完成针对地名、人名和时间这三种实体的识别模型的构建。

条件随机场模型(Conditional Random Fields, CRFs)是2001年Lafferty等人提出的[12]。它由隐马尔科夫模型(Hidden Mark · OV Model,HMM)、最大熵马尔科夫模型(Maximum Entropy Markov Model,MEMM)逐级发展而来,并克服了两者的缺陷,主要应用于基于条件概率处理序列标注问题。CRFs是一种判别式模型,采用的是无向图分布,没有严格的独立性假设,因此可以任意选取特征,如图1所示。

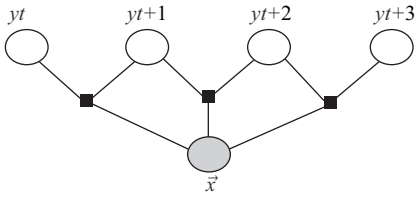

图1　条件随机场模型

在条件随机场中,一串y对一串x,将多个x作为一个整体来确定y和y之间的转移概率。例如,在本文所使用的经过人工标注的先秦典籍实体语料中,x可以表示一个被标注的实体,y则表示该实体中每一个分布序列。在上述对条件随机场模型分析的基础上,条件随机场的公式表示为:

$$y' = \underset{y}{\operatorname{argmax}} p_{\vec{\lambda}}(\vec{y} \mid \vec{x}) = \underset{y}{\operatorname{argmax}} \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \exp$$

$$\left(\sum_{j=1}^{n} \sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j) \right) \qquad \text{(公式1)}$$

其中,$p_\lambda(Y|X)$是在λ参数制约下类别向量Y的条件概率,$Z(x)$是归一化因子,n是序列长度,m是特征函数个数。特征函数有两种:一是状态特征函数,二是转移特征函数。基于这两个函数,条件随机场模型不仅能够使用实体自身的特征,而且还可以增加外在的特征,从而可以有效地提升模型的整体性能。因此,本文将依照条件随机场模型构建自动识别模型,并将其应用于先秦语料库中人名、地名、时间实体的识别。

3　人名、地名、时间实体的内部数量统计和外部特征分析

3.1　三类实体的内部数量统计

在上述规模为465,197个词汇的先秦语料库中,5300个人名共出现24,615次,出现次数最多的“晏子”有968个,这些人名实体占词汇总数的5.29%;地名共出现14,424次,占词汇总数的3.10%,1666种地名中“晉”出现了1463次;261种时间词共出现6042次,占词汇总数的1.30%,其中“今”出现715次。三类实体的词汇共占整个语料库词汇的9.70%,从实体这一语言单位的整体数量上来看,具有一定的代表性,识别这三类实体对于探究历史事件的构成具有一定的价值和意义。这三类实体的词长、频数、所占本类实体比重如表2所示。

由表2中人名、地名、时间实体的长度情况可知,我们需要识别的实体长度主要为2左右,如“晏子”“齊侯”“齊國”“邾婁”“正月”“元年”等。类似“公孫鍾離”“長狄緣斯”“十有二月”这些长度超过3的实体较少出现,实体长度最长为5,因此我们选择特征窗口为5,范围是{-2,-1,0,1,2}。

表2　三类实体的词长、频数、比重

人名			地名			时间词		
词长	频数	比重	词长	频数	比重	词长	频数	比重
2	16382	66.55%	1	11685	81.01%	2	3857	63.84%
3	4431	18.00%	2	2684	18.61%	1	1333	22.06%
1	2969	12.06%	3	37	0.26%	4	443	7.33%
4	818	3.32%	4	17	0.12%	3	324	5.36%
5	15	0.06%	5	1	–	5	85	1.41%

3.2　三类实体的外部特征分析

语言中的句子是以一个线性序列呈现出来的，若将语言的句子序列表示为“SL_m，…，SL_i…，SL_1，[R，R_1，…，R_n]，SR_1，…，SR_j，…，SR_n”，把 SL_1 和 SR_1 界定为事件基本实体构件的一元左右边界词，SL_2、SL_1 和 SR_1、SR_2 界定为二元左右边界词，SL_3、SL_2、SL_1 和 SR_1、SR_2、SR_3 界定为三元左右边界词，由于事件的基本实体构件人名、地名、时间实体与上下文有关，所以需要关注这三类实体左右边界的词汇知识分布状况。

(1)左右边界词词类特征

考虑到部分一元边界词的词类本身在语料库中出现频率较高，将其统计为人名、地名、时间实体左右边界词词类特征时不具较好的特征性，因此本文所研究的先秦语料库中的事件实体一元左右边界词词类分布由下述公式获取：

$$\text{边界词词类频率}=\frac{\text{该词类作为一元边界词出现的次数}}{\text{该词类在语料库中出现的次数}}\times 100\% \quad \text{(公式 2)}$$

综合考虑人名、地名、时间这三类实体的左右边界词的词类分布特征，按照公式 2 进行计算。为了将来更好地推广和应用所构建的实体识别模型，本文基于 Python 语言搭建了一个面向先秦典籍的实体识别平台，该平台由特征统计、CRF 训练、效果评价和模型应用四部分主要功能构成。这一平台的功能结合本文探究的不同步骤逐一展现出来。三类实体的左右边界词词类分布状况统计功能如图 2 和图 3 所示。

图2　特征统计功能截图(左边界词性)

图3　特征统计功能截图(右边界词性)

基于上述平台的统计功能截图可知,介词、使动用法、为动用法、意动用法在三类实体的边界词中较有特点。具体考虑三种事件基本实体构件中每一类实体的左右高频边界词词类,统计结果见表3—5。

表3　人名实体左右高频边界词词类分布

编号	左边界词类			编号	右边界词类		
	标注符号	词性	频率		标注符号	词性	频率
1	w	标点	58.93%	6	v	动词	50.11%
2	v	动词	20.81%	7	w	标点	24.06%
3	n	普通名词	5.14%	8	u	助词	5.93%
4	ns	地名	4.88%	9	d	副词	5.11%
5	p	介词	3.66%	10	p	介词	3.85%

表4　地名实体左右高频边界词词类分布

编号	左边界词类			编号	右边界词类		
	标注符号	词性	频率		标注符号	词性	频率
1	v	动词	38.01%	6	w	标点	48.95%
2	w	标点	33.05%	7	n	普通名词	12.94%
3	p	介词	22.15%	8	v	动词	9.76%
4	c	连词	2.52%	9	nr	人名	8.33%
5	r	代词	1.06%	10	u	助词	7.09%

表5　时间实体左右高频边界词词类分布

编号	左边界词类			编号	右边界词类		
	标注符号	词性	频率		标注符号	词性	频率
1	w	标点	50.86%	6	w	标点	47.62%
2	t	时间词	20.62%	7	t	时间词	20.62%
3	n	普通名词	18.12%	8	n	普通名词	15.23%
4	nr	人名	5.59%	9	v	动词	4.27%
5	v	动词	1.54%	10	r	代词	3.13%

由表3—5可知,人名、地名、时间实体的一元左右边界词的词类主要是集中在标点、动词、普通名词上,这三种词的特征知识应该重点考虑。因此,在通过条件随机场训练人名、地名、时间实体的过程中,考虑词性特征知识时不仅要考虑将使动用法、为动用法、意动用法加入到特征模板中去,还要考虑三类实体分别的左右边界词词类特征。

(2)各实体左右边界词特征

若命名实体的左右边界词确定了,则这个实体可以被识别出,因此可以将三类实体具体的左右边界词作为特征之一加入到特征模板中。本文对人名、地名、时间实体的左右边界词频次分别进行统计①,并将其中前10个高频词列出,具体如表6—8所示。

表6　人名实体左右边界词及其数量

编号	左边界词	数量	右边界词	数量
1	。	4886	、	2386
2	,	3730	曰	2226
3	、	2008	,	1602
4	”	1093	之	1449
5	」	1034	。	1404
6	使	654	以	466
7	於	448	卒	466
8	“	400	使	409
9	與	327	來	391
10	晉	293	如	379

表7　地名实体左右边界词及其数量

编号	左边界词	数量	右边界词	数量
1	于	1862	。	3303
2	,	1406	,	2861
3	。	1340	之	993
4	、	1271	、	693
5	伐	902	人	395
6	於	736	師	361
7	如	542	也	316
8	自	290	而	189
9	及	272	侯	130
10	奔	233	以	126

表8　时间实体左右边界词及其数量

编号	左边界词	数量	右边界词	数量
1	。	2212	,	2661
2	,	403	春	648
3	王	299	。	164
4	夏	288	之	102
5	秋	272	朔	81
6	冬	254	君	73
7	”	122	者	60
8	八月	100	而	43
9	六月	88	吾	35
10	五月	86	不	35

① 在本文中,把标点符号也看作词汇,因此统计的高频词中有标点符号存在。

结合三种实体所共有的左右高频词分布情况,“。”“,”“以”“、”“曰”“之”“于”“於”“伐”“使”“也”等,这些具有语义区别的标点、虚词可被提取出作为特征,将对识别三类实体有较大帮助。

4 先秦语料的预处理

在对语料进行人名、地名、时间实体自动识别前,需要对标注好的语料进行预处理。首先,去除先秦语料中人名、地名、时间实体以外的其他结构的标记;其次,针对三种实体的具体特征,确定标注集。在确定标注集的过程中,主要参考下面这一公式:

$$L_K = \frac{1}{N}\sum_{i=j}^{k} iN_i \qquad (公式3)$$

上式中,L_k 表示当 $i \leqslant k$ 时,人名、地名、时间实体的平均加权后的长度,N_i 表示先秦语料库中长度为 i 的事件基本实体构件出现的次数,k 与 j 分别表示语料库中最长和最短事件基本实体构件的长度,N 表示语料库中人名、地名、时间实体分别出现的次数。基于公式3,结合先秦语料的基本情况以及相应的实验结果,识别模型构建中确定使用4词位的标注集。

根据具体不同标记长度的知识抽取实验,在模型构建中所用的词位标注集由十三个不同的标记组成,标注集用T来表示,具体为T={BN,MN,EN,SN,BP,MP,EP,SP,BT,MT,ET,ST,A},以B、M、E开头分别表示实体的初始词、实体中间词、实体结束词;S开头表示该实体是单字结构;以N,P,T结尾分别表示是人名、地名、时间实体;A代表需要识别的三类实体以外的词。例如:BN表示人名初始词,MN表示人名中间词,EN表示人名结束词,SN表示单字人名。经过以上的处理,得到的部分语料见表9。

表9 先秦语料预处理结果样例

序号	词语	词性	标记
1	夏	n	A
2	四	t0	BT
3	月	t1	ET
4	,	w	A
5	取	v	A
6	邬	ns	SP
7	大	n0	A
8	鼎	n1	A
9	于	p	A
10	宋	ns	SP
11	。	w	A

5 模型构建流程和效果评价方法

先秦历史事件基本实体构件自动识别的语料模型在条件随机场模型构建的基础上由三个环节构成。第一,训练语料获得模板。我们基于第3部分寻找的语料库特征制作特征模板,将其按照9:1、8:2、7:3、6:4的比例分别处理为CRFs训练和测试文本,利用CRF++ 0.58进行训练[13]。第二,测试事件基本实体构件自动识别效果。测试训练结果是在第一步训练得到的特征模板的基础上,对测试集自动抽取人名、地名、时间实体的过程。第三,模型构建效果评价。训练所得模板的效果必须要通过一定评价指标来量化。

5.1 训练文本特征添加及处理

CRFs本身的一个突出优点是可以任意加入与处理对象有关的语言学特征,通过统计先秦语料库中历史事件基本构件的三个组成部分——人名、地名、时间实体,发现其中一些特征知识可以加入到特征模板中,具体特征知识如下。

(1)词长分布特征。先秦语料库三类构成事件基本构件的词中,长度为2的词最多,如:又娶二女於【戎】,大戎【狐姬】生【重耳】,【小戎子】生【夷吾】。词长分布可以作为特征之一加入到特征模板中。

(2)左右边界词词类特征。副词、介词、为动用法、意动用法在第3部分三类实体外部特征统计中出现次数相对较多。若属于这四类词,标注为“Y”,否则标为“N”。

(3)具体左右边界词特征。先秦语料中出现次数较多的“,”“以”“、”“曰”“之”“于”“於”“伐”“使”“也”可以作为特征加入到特征模板中。确定了边界词也就确定了实体,因此我们将这十个词作为特征加入,将这些词标注为“Y”,否则标为“N”。

按照上述特征标注,将得到如表10的条件随机场训练文本。

表10　先秦语料的训练文本样例

序号	词语	是否边界词词性	是否边界词	标记
1	夏	N	N	A
2	四	N	N	BT
3	月	N	N	ET
4	,	Y	Y	A
5	取	N	N	A
6	鄙	N	N	SP
7	大	N	N	A
8	鼎	N	N	A
9	于	Y	Y	A
10	宋	N	N	SP
11	。	N	N	A

5.2　评价指标选择

本文选取准确率P(Precision)、召回率R(Recall)这两个广泛应用于信息检索和统计学分类中的度量值来评价模型构建结果的质量。公式如下:

$$\text{准确率 P}=\frac{\text{正确识别的实体}}{\text{正确识别的实体+被错误识别的实体}}\times 100\% \quad (\text{公式 4})$$

$$\text{召回率 R}=\frac{\text{正确识别的实体}}{\text{正确识别的实体+未被识别的实体}}\times 100\% \quad (\text{公式 5})$$

准确率和召回率是互相影响的,理想情况下是两者都高,但通常情况下是准确率提高,召回率会下降,反之亦然。在准确率和召回率出现相互矛盾的情况下,需要综合考虑两者,这里引入调和平均数F值(F-Measure)对准确率和召回率进行加权调和平均,如公式6,选择$\beta=1$。

$$\text{调和平均值 F}=\frac{(\beta^2+1)\times P\times R}{(\beta^2\times P)+R}=\frac{2\times P\times R}{P+R} \quad (\text{当}\beta=1) \quad (\text{公式 6})$$

5.3　人名、地名、时间实体自动识别模型的测评过程

为取得更好的测试结果,我们采取交叉验证的方法,将语料以句子为单元,全部打乱,均分成十份,并按照9:1、8:2、7:3、6:4的比例分别分成训练集和测试集,以期从中获得最科学合理的自动抽取模型。平台具体的模型训练功能如图4所示。

图4　CRFs训练测试功能截图

先将先秦语料处理为可用于条件随机场训练的文本,可以选择处理为带特征文本或不含特征文本,进一步选择不同的训练和测试比例,该可视化软件可以自动将文本按比例分成训练集和测试集。点击“训练”,软件将执行训练指令并生成模板;测试是在生成模板的基础上,对测试集自动进行人名、地名、时间实体的自动抽取。在这里我们将按照9:1、8:2、7:3、6:4的训练测试比,分成含特征的训练与测试文本,分别进行训练。

实体识别效果评价功能如图5所示,选择要进行条件随机场自动抽取实体效果评价的文本,由于在训练模型时我们选择的是带特征文本,在这里我们点击“带特征”,该可视化软件将自动对其中机器基于特征模型抽取出的三类实体与人工标注的三类实体进行对比,并利用上文5.2部分选择的评价指标进行正确率、召回率、调和平均数的计算。

图5　CRFs测试效果评价功能

按照模型构建流程,最终得到按7:3的比例分割整体获得的模板测试效果最好,这里只展示7:3比例分割后测试的结果,并对人名、地名、时间实体识别效果进行加权计算。具体每份模型的测试结果的准确率、召回率、F值见表11。

表11　7:3比例分割训练测试效果汇总

序号	训练语料	测试语料	准确率(P)	召回率(R)	F值(F)
1	1—7	8—10	90.35%	84.50%	87.32%
2	2—8	1,9,10	90.25%	84.36%	87.20%
3	3—9	1,2,10	90.27%	84.64%	87.36%
4	4—10	1—3	90.58%	84.39%	87.37%
5	1,5—10	2—4	90.48%	84.00%	87.12%
6	1,2,6-10	3—5	90.38%	84.07%	87.11%
7	1—3,7—10	4—6	90.21%	83.77%	86.86%
8	1—4,8—10	5—7	89.85%	83.21%	86.39%
9	1—5,9,10	6—8	90.41%	83.73%	86.94%
10	1—6,10	7—9	90.10%	83.76%	86.80%

通过对比十次测试结果,我们可以看出训练语料的不同选择对最终测试结果有一定影响。序号 4 的综合评价指标最佳,故可以确定 4 号语料获得的模型为先秦典籍历史事件基本实体构件自动识别的最佳模型。

5.4 人名、地名、时间实体自动识别错误分析

对自动识别过程中产生的错误进行原因分析,发现出错原因主要包括以下四个方面。

第一,低频率出现的单字事件基本实体构件难以被识别。例如,“遂殺其二子【幕/nr】及平夏”中,“幕”是人名,但是在整个语料中出现次数较少,没有被自动识别出来;又如“且【鱄/nr】實使之”,“鱄”作为人名在训练文本中一共出现两次,在机器自动抽取实体时,也没有将其识别出。

第二,左右边界词标注令机器识别出错。如“使治亂存亡若【高山/ns】之與深谿,若白垩之與黑漆,則無所用智,雖愚猶可矣”,“高山”不是地名,是普通名词,在这里被错误地识别。“高山”前的“若”在处理为含有特征知识的文本时被处理为“若#N#Y#A”,“高山”后的“之”被处理为“之#Y#Y#A”,“高山”左右含有边界词特征,因此在这里被错误地抽取出。

第三,一词多义造成识别错误。例如,“楚【司馬/nr】子庚聘于秦”,“司馬”在此处是官职名称,但“司馬”还可以是人的姓氏,自动识别时,机器将“司馬子庚”识别为人名实体,造成错误。

第四,人工标注不合法,存在误标和漏标。例如,“自今以來,【亶父/nr】非寡人之有也,子之有也”,“亶父”是人名实体,但在人工标注时,误标为地名实体。标注前后不一致,如“鄭伯”中“鄭”被标为“ns”,但“鄭伯”被标注为“nr”。

但从整体效果上看,对于出现频率高的事件基本实体构件的识别效果较为突出,比如“晏子”“晉侯”“楚”“昔”。

5.5 先秦历史事件基本实体构件自动识别模型的应用

在本文获得的序号 4 的模型基础上,可以应用该模型对其他历史典籍快速且较为准确地识别其中的人名、地名、时间实体。本文利用开发的软件平台,将序号为 4 的模型应用于同为先秦时期的《国语》语料的人名、地名、时间实体自动识别上,具体的平台功能如图 6 所示。

图 6 《国语》实体自动抽取效果

将自动抽取的《国语》中三类实体与人工标注进行对比,得到该模型在《国语》上的识别已可以达到81.29%的调和平均值。将机器自动标注结果还原为原文章,得到图7的结果。

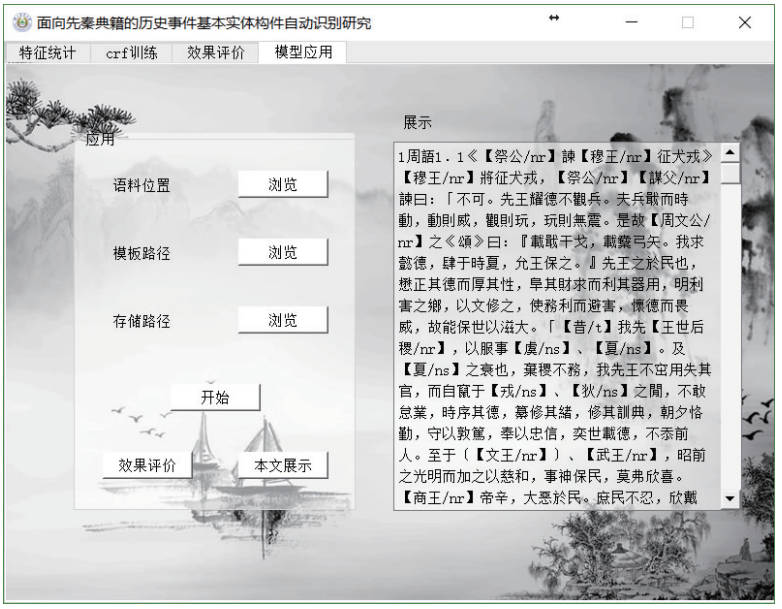

图7 《国语》自动抽取实体展示

(1)《国语》自动抽取实体成果展示

《国语》是一部国别体史书,以国家为单位,分别记叙历史事件,因此对整本《国语》统计三类实体的抽取成果意义不大。本文对《国语·越语》中被自动识别出的实体进行分析,由此窥见该部分记述的历史事件,并利用Python绘制词云进行可视化展现,见图8。

图8 《国语·越语》实体抽取词云展示

Python生成词云的本质是对文本中的词进行统计,根据出现频率的多少来按比例展示大小。我们将模型自动识别出的三类实体抽取出,制作出该词云。由图8可以明显看出,《国语·越语》中对“范蠡”“勾践”“夫差”三位人物的描述较多;事件发生地点主要集中在“会稽”和“越国”;时间实体出现较少,偶尔出现“三月”“四年”“五年”这类表示时间的词。通过词云展示所识别的实体,可以有效展现历史事件中的基本实体。

(2)模型应用的价值与局限

机器可以基于模型自动识别出大部分人名、地名、时间实体,整体性能较为突出,本文选取自动识别的《国语》语料中“邵公”所言的几个句子具体分析该模型的使用价值与局限。

【厲王/nr】虐,國人謗王。【邵公/nr】告曰:「民不堪命矣!」王怒,得【衛/ns】巫,使監謗者,以告,則殺之。國人莫敢言,道路以目。王喜,告【邵公/nr】曰:「吾能弭謗矣,乃不敢言。」【邵公/nr】曰:「是障之也。防民之口,甚於【防川/ns】。

上面句子中,“衛”作为地名在训练语料中出现次数较多,在此被正确识别出;在训练语料中出现非常少的“厲王”“邵公”,由于其与训练集中部分人名有相同结构,因此也被识别出;与此同时,“防川”不是地名,但是其前的“於”具有识别三类实体的边界词特征,因此在此处被错误地识别成地名实体。

【邵公/nr】曰:「【昔/t】吾驟諫王,王不從,是以及此難。【今/t】殺【王子/nr】,王其以我為懟而怒乎!夫事君者險而不懟,怨而不怒,況事王乎?」乃以其子【代宣王/nr】,【宣王/nr】長而立之。

“【今/t】殺【王子/nr】”一句中“王子”被错误识别出,一是由于受训练集中“晏子”等人名结构的影响,二是前“殺”字的影响。本段中,第一个“宣王”被错误识别为“代宣王”,第二个则被正确识别,原因在于上下文语义的影响。

【邵公/nr】以告【單襄公/nr】曰:「【王叔子/nr】譽【溫季/nr】,以為必相【晉國/ns】,相【晉國/ns】,必大得諸侯,勸二三君子必先導焉,可以樹。【今/t】夫子見我,以【晉國/ns】之克也,為己實謀之。

上段中并未出现识别错误,所识别的词语也是常见的人名、地名和时间实体,效果比较理想。总之,在将模型应用于其他语料的事件基本实体构件自动识别上时,具有一定的局限性,只有当被识别的语料与训练模型的语料相似时,才能达到令人满意的应用效果。

6 结语

人名、地名、时间实体作为事件基本的实体构件,对掌握历史事件起着重要作用。本文对这三类实体进行内部数量统计和外部特征分析,构建了面向先秦典籍历史事件基本实体构件自动识别的特征模板。与此同时,应用条件随机场模型,对先秦语料库语料进行交叉验证,在9:1、8:2、7:3、6:4不同的训练测试比中选取调和平均数最高(87.37%)的模型作为古汉语历史事件基本实体构件自动识别的最佳模型。最后将该模型应用到《国语》语料的事件基本实体构件自动识别上,取得调和平均数为81.29%的效果,证明该模型基本达到了实用的目标。在后续研究中,将增加训练语料及融入更多特征以改进现有模型的精确度与召回率,进一步推广该模型在古汉语语料上的使用。

参考文献

1 刘炜,等.面向人文研究的国家数据基础设施建设[J].中国图书馆学报,2016(5):29-39.

2 欧阳剑.面向数字人文研究的大规模古籍文本可视化分析与挖掘[J].中国图书馆学报,2016(2):66-80.

3 Unsworth, J. What Is Humanities Computing and What Is Noz[EB/OL].[2017-03-26].http://computerphilologie.uni-muenchen.de/jg02/unsworth.html.

4 柯平,宫平.数字人文研究演化路径与热点领域分析[J].中国图书馆学报,2016(6):13-30.

5 赵妍妍,等.中文事件抽取中事件类别的自动

识别[C]//孙茂松.第三届学生计算语言学研讨会论文集.沈阳:东北大学出版社,2006:240-245.

6 徐元子,等.基于条件随机场的网络评论与事件中命名实体匹配研究[J].计算机应用研究,2016(6):1642-1647.

7 陈箫箫,刘波.微博中的开放域事件抽取[J].计算机应用与软件,2016(8):18-22,109.

8 陈慧炜.刑事案件文本信息抽取研究[D].南京:南京师范大学,2011.

9 景悦诚.基于丰富语言特征的中文社交媒体事件发掘[D].上海:上海交通大学,2015.

10 李劲,等.面向个人简历的事件抽取和检索框架[J].计算机科学,2012(7):154-160.

11 黄水清,等.基于先秦语料库的古汉语地名自动识别模型构建研究[J].图书情报工作,2015(12):135-140.

12 Lafferty J, et al. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C].Proceedings of 18th International Conference on Machine Learning.2001:282-289.

13 CRF++[EB/OL].[2017-04-21].http://crfpp.sourceforge.net/.

(王东波 副教授 南京农业大学信息科学技术学院,高瑞卿 南京农业大学信息科学技术学院信息管理与信息系统专业2015级本科生,沈 思 讲师 南京理工大学经济管理学院,李 斌 副教授 南京师范大学文学院)

收稿日期:2017-08-28