文章编号:1000－2367(2015)04－0163－05 DOI:10.16366/j.cnki.1000－2367.2015.04.030

体育赛事命名实体识别研究

谷 川[a],宋 旭[b]

(安阳师范学院 a.软件学院;b.计算机与信息工程学院,河南 安阳 455000)

摘 要:为了准确地从中文文本中识别出复杂体育赛事命名实体,提出了一种基于双层条件随机场模型的命名实体识别方法.该方法首先在低层条件随机场模型中识别出简单体育赛事命名实体,然后在高层条件随机场模型中识别出嵌套了简单体育赛事命名实体的复杂命名实体如赛事名、参赛球队名和比赛场馆名.在对大规模真实语料进行的开放测试中,赛事名、参赛球队名和比赛场馆名识别的 F 值分别达到 97.09%, 97.81% 和 98.03%.

关键词:命名实体识别;体育赛事领域;双层条件随机场

中图分类号:TP391.11 **文献标志码**:A

随着社会文明的进步发展,体育在现代生活中所占的比重与日俱增,人们对体育赛事的关注程度越来越高.目前,互联网上每天出现的体育赛事信息层出不穷,如何从海量的体育赛事信息中抽取出人们感兴趣的内容是亟待解决的问题.而体育赛事命名实体识别是体育赛事信息抽取的基础,所以研究体育赛事命名实体识别显得非常重要.

体育赛事命名实体是指对理解体育新闻起决定作用的赛事名、赛事级别、比赛场馆名、参赛球队名、运动员姓名、比赛时间和比赛结果这 7 类实体.目前,关于中文命名实体识别研究虽然已经很多[1－6],但是针对体育赛事命名实体识别的研究却很少.文献[7]使用隐马尔科夫模型与规则相结合的方法进行体育赛事实体识别,而隐马尔科夫模型是一种生成式模型,为保证推导的正确性,需要做出严格的独立性假设.条件随机场(Conditional random fields,CRF)是 Lafferty 等人于 2001 年在隐马尔科夫模型和最大熵模型的基础上提出的一种用于序列标注的条件概率模型,能方便地在模型中包含多种特征,它不需做出要严格的独立性假设,又较好地解决了标注偏置的问题.文献[8]显示,利用条件随机场对体育赛事实体进行识别,取得了较好识别效果,而对于赛事名、参赛球队名和比赛场馆名这三类实体的识别效果却不理想,原因是它们的构成比较复杂,经常会在实体内部包含赛事名简称、球队名简称和简单地名这些简单命名实体.本文针对这一问题提出了一种利用双层条件随机场进行体育赛事命名实体识别的方法,该方法先在低层进行赛事名简称、球队名简称和简单地名的识别,然后将识别结果传递到高层模型,由高层模型识别出赛事名、参赛球队名和比赛场馆名.实验结果表明,该方法对于以上三类复杂命名实体的识别取得了很好的效果.

1 条件随机场

CRF 是一种以给定的输入结点值为条件来预测输出结点值概率的无向图模型[9],它通过定义标注序列和观察序列的条件概率 $P(Y|X)$ 来预测最可能的标注序列.命名实体识别过程就是一个序列标注过程[10].因此,CRF 非常适合用于命名实体识别.最简单最常用的 CRF 是线性结构的 CRF,如图 1 所示.

用 $X=(x_1,x_2,\cdots,x_T)$ 表示一个句子的字词序列即观察序列,$Y=(y_1,y_2,\cdots,y_T)$ 表示一个句子对应的标注符号序列即标注序列.在给定观察序列 X 的条件下,标注序列 Y 的条件概率分布 $P(Y|X)$ 构成 CRF 模型,公式如下:

收稿日期:2014-09-24;**修回日期**:2015-03-17.

基金项目:国家自然科学基金(60875081);河南省基础与前沿技术研究计划项目(112300410182).

作者简介(通信作者):谷 川(1980－),男,河南郾城人,安阳师范学院讲师,研究方向为自然语言处理与机器学习,E-mail:jxk-20@163.com.

$$P(Y \mid X) = \frac{1}{Z_0}\exp\Big[\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(y_{t-1}, y_t, x, t)\Big], \tag{1}$$

其中，$Z_0 = \sum_{Y}\exp\Big[\sum_{t=1}^{T}\sum_{k=1}^{K}\lambda_k f_k(y_{t-1}, y_t, x, t)\Big]$是归一化因子；$f_k(y_{t-1}, y_t, x, t)$ 是一个二值的状态特征函数，λ_k 是特征函数 $f_k(y_{t-1}, y_t, x, t)$ 的权重，通过训练得到[14].

对于观察序列 X，最佳的标注序列 Y 满足如下公式：

$$Y = \underset{Y}{\operatorname{argmax}} P(Y \mid X). \tag{2}$$

2 基于双层条件随机场的体育赛事命名实体识别

2.1 双层条件随机场模型

赛事名、参赛球队名和比赛场馆名构成比较复杂，经常出现嵌套现象，即实体内部中会包含赛事名简称、球队名简称和简单地名，如："2013 赛季亚冠联赛"、"广州恒大"、"北京工人体育场"等. 所以更好的办法是将这些复杂命名实体的识别放在简单命名实体被识别之后进行. 本文通过构建双层 CRF 模型，将复杂命名实体的识别分成两步进行，首先在低层模型中完成对用于对赛事名简称(如"亚冠")、球队名简称(如"恒大"或"恒大队")和简单地名(如"广州"和"北京")等这些简单命名实体的识别，然后将低层识别的结果传递到高层模型，这样在高层模型中既有观察值，又有低层 CRF 模型的识别结果，从而提高了复杂命名实体的识别效果. 双层 CRF 模型如图 2 所示.

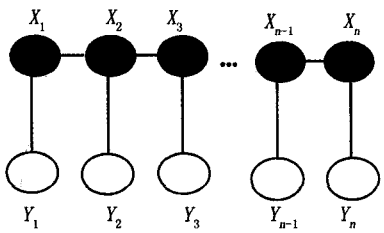

图 1 线性结构的CRF模型

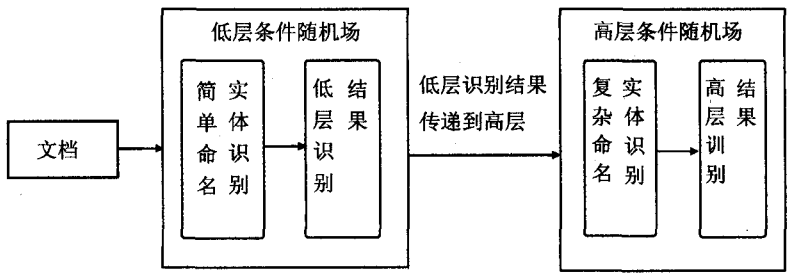

图 2 基于双层条件随机场模型的识别框架

2.2 低层条件随机场模型

2.2.1 训练语料的预处理

通过仔细分析赛事名简称、球队名简称和简单地名的词法特点，发现这三类实体文本内部没有太多的联系，并且字与字之间成词规律不明显，所以选择在字一级粒度上进行建模来实现对这些实体的识别.

针对字粒度的赛事名简称、球队名简称和简单地名的识别任务，定义了包含 7 种标记的集合 $L=\{EB, EI, TB, TI, LB, LI, O\}$，其中各标记分别代表赛事名简称开始、赛事名简称内部、球队名简称开始、球队名简称内部、地名开始、地名内部、其他. 例如下列语句"2013 赛季中超联赛第 29 轮比赛全面打响，杭州绿城队坐镇主场迎来贵州人和队的挑战"，采用以上定义的标记集标注结果为："2/O 0/O 1/O 3/O 赛/O 季/O 中/EB 超/EI 联/O 赛/O 第/O 2/O 9/O 轮/O 比/O 赛/O 全/O 面/O 打/O 响/O，/O 杭/LB 州/LI 绿/TB 城/TI 队/TI 坐/O 镇/O 主/O 场/O 迎/O 来/O 贵/LB 州/LI 人/TB 和/TI 队/TI 的/O 挑/O 战/O".

2.2.2 特征模板的设定

特征模板可以看作是对一组上下文的特征按照共同属性进行的抽象[11]. 为了建立反映语言内在规律的条件随机场模型，需要通过定义模板来筛选特征[12]. 条件随机场很容易将多种上下文特征融合到模型中. 本文以赛事名简称的特征模板为例讲述低层模型中模板的设定过程.

赛事名简称的前缀多为洲际名或国家名的简称，如"世"、"亚"、"澳"等，它们对于赛事名简称左边界的识别起了很大作用，可以把这些常用前缀作为可利用的特征. 一些字出现赛事名简称的中的概率很大，如"会"、"杯"、"甲"等，这些常用字也是实体识别的有用信息. 另外，指界词对确定赛事名简称的边界起着重要作用，根据出现位置的不同，分为左指界词和右指界词，如"赛季"，"联赛"等，这些词是识别赛事名简称的非常重要

上下文特征.

根据以上考虑，经过反复比较实验，选定的原子特征模板如表 1 所示. 为了描述和表示上下文特征，本文把上下文窗口的大小设定为 5. 球队名简称和简单地名的特征模板与赛事名简称的特征模板类似，这里不再详述.

表 1　赛事名简称的原子特征模板

序号	原子模板	模板含义
1	CurChar	当前字
2	EventNameShortPrefix	当前字是否为赛事名简称前缀
3	EventCommonChar	当前字是否为赛事名简称常用字
4	LeftBoundary	当前字前面两个字中是否含赛事名简称的左指界词
5	RightBoundary	当前字后面两个字中是否含赛事名简称的右指界词

2.3　高层条件随机场模型

2.3.1　训练语料的预处理

在高层条件随机场模型中，主要是识别赛事名、参赛球队名和比赛场馆名. 这些复杂的命名实体是由赛事名简称、球队名简称、简单地名及其他一些词构成，而词内部联系紧密，如果把词拆分为单字进行识别将严重损失文本的特征信息. 文本中词性的前后依赖关系对命名实体的识别具有启发作用. 因此，在高层模型中以词为粒度进行建模，同时需要把词性特征加入到 CRF 模型中.

针对词粒度的复杂命名实体识别任务，仍采用低层模型中定义的标注集，只是 5 种标记的含义发生了变化，它们依次表示赛事名开始(EB)、赛事名内部(EI)、参赛球队名开始(TB)、参赛球队名内部(TI)、比赛场馆名开始(LB)、比赛场馆名内部(LI)、其他(O). 对训练语料预处理时，首先使用中科院计算所开发的汉语词法分析系统 ICTCALS 对原始语料进行分词和词性标注，然后将低层模型中识别出来的赛事名简称对应的词串标注为/ne、球队名简称对应的词串标注为/nt、简单地名对应的词串标注为/ns. 例如下列语句“北京时间 10 月 30 日 19:35，2013 赛季中超联赛第 29 轮 8 场比赛同时打响，大连阿尔滨坐镇金州体育场迎战辽宁宏运.”，经过分词和词性标注后转变为“北京/n 时间/n 10 月/t 30 日/t 19:35/m，/w 2013/m 赛季/n 中/f 超/v 联赛/n 第 29/m 轮/q 8/a 场/q 比赛/v 同时/c 打响/v，/w 大连/n 阿尔/n 滨/g 坐镇/v 金州/n 体育场/n 迎战/v 辽宁/n 宏/n 运/v ./w”. 一方面在低层模型中，“中超”被识别为赛事名简称，“阿尔滨”和“宏运”被识别为球队名简称，“北京”和“金州”被识别为简单地名，另一方面“2013 赛季中超联赛”整体是一个赛事名，“大连阿尔滨”和“辽宁宏运”是两个参赛球队名，“金州体育场”是一个比赛场馆名. 因此，采用已定义的标注集将该语句转换为“北京/ns/O 时间/n/O 10 月/t/O 30 日/t/O 19:35/m/O，/w/O 2013/m/EB 赛季/n/EI 中超/ne/EI 联赛/n/EI 第 29/m/O 轮/q/O 8/a/O 场/q/O 比赛/v/O 同时/c/O 打响/v/O，/w/O 大连/ns/TB 阿尔滨/nt/TI 坐镇/v/O 金州/ns/LB 体育场/n/LI 迎战/v/O 辽宁/ns/TB 宏运/nt/TI. /w/O”.

2.3.2　特征模板的设定

对于条件随机场而言，特征模板的设定非常关键. 本文以赛事名的特征模板为例讲述高层模型中模板的设定过程. 在赛事名中通常会包含赛事名简称和简单地名，而赛事名简称和简单地名已经在低层模型中被识别出来，因此这两种实体就是赛事名识别可以利用的资源. 通过分析赛事名的构成，发现以下信息：首先，赛事名通常以年份数字如“2013”或“2013/14”开头、以特征词如“联赛”或“常规赛”结尾，所以年份数字和特征词可以作为识别赛事名的前缀和后缀特征. 其次，赛事名通常为名词性短语，词性是一个明显的识别特征. 最后，与赛事名相邻的指界词对于赛事名的出现具有指示作用，所以指界词也是识别实体的明显特征. 综合以上考虑，在多次实验基础上得到了原子特征模板如表 2 所示，这里把上下文窗口的大小也设置为 5. 除了原子特征模板，为了描述更加复杂的语言现象，还需要设定复合特征模板. 复合特征模板如表 3 所示. 参赛球队名与比赛场馆名的特征模板和赛事名的特征模板基本类似，这里不再详述.

3　实验与分析

3.1　实验数据

为了客观的评价双层条件随机场模型对赛事名、参赛球队名和比赛场馆名的识别效果，本文设计了 4 次

开放测试实验:实验 1 采用基于隐马尔科夫模型的方法;实验 2 采用隐马尔科夫模型与规则相结合的方法;实验 3 采用基于单层条件随机场模型的方法;实验 4 采用基于双层条件随机场模型的方法.这 4 次实验所采用的训练语料和测试语料完全相同.从新浪体育和搜狐体育这两个专业网站下载有关体育赛事的网页 1200 个.从中任意抽取 400 个网页作为开放测试语料,其余 800 个网页作为训练语料,如表 4 所示.

表 2 赛事名的原子特征模板

序号	原子模板	模板含义
1	CurChar	当前
2	CurWordPos	当前词词性
3	EventNameShort	当前词是否为赛事名简称
4	LocationName	当前词是否为简单地名
5	YearNumber	当前词是否为年份数字
6	EventNameFeature	当前词是否为赛事名特征词
7	LeftBoundary	当前词是否为赛事名左指界词
8	RightBoundary	当前词是否为赛事名右指界词

表 3 赛事名的复合特征模板

序号	复合特征模板
1	CurWord(n) & CurWord($n+1$) ($n=-1,0$)
2	CurWord(n) & CurWordPos(n) ($n=-2,-1,0,1,2$)
3	EventNameShort(n) & EventNameFeature($n+1$) (n=0)
4	LeftBoundary($n-1$) & AfterFeature(n) ($n=0$)
5	EventNameFeature(n) & RightBoundary($n+1$) (n=0)

表 4 语料信息

语料名称	赛事名个数	参赛球队名个数	赛场馆名个数
训练语料	800	5243	738
测试语料	400	2681	356

本文采用准确率(P)、召回率(R)和综合指标 F 值作为系统识别效果的评价指标.其中,准确率衡量的是系统的查准率,召回率衡量的是系统的查全率,它们的计算公式定义如下:

$$P=\frac{\text{正确识别的命名实体数}}{\text{识别出的命名实体总数}}\times 100\%, \tag{3}$$

$$R=\frac{\text{正确识别的命名实体数}}{\text{文本中的命名实体总数}}\times 100\%, \tag{4}$$

准确率和召回率有时会出现不一致的情况,即准确率高时,召回率低,反之亦然.评估一个系统的整体性能时,就需要综合考虑它们,最常采用的办法就是 F 值,计算公式如下:

$$F=\frac{(\beta^2+1)\times P\times R}{(\beta^2\times P+R)}\times 100\%, \tag{5}$$

其中,β 是 P 和 R 之间权衡因子,在这里 β 取值为 1,即认为 P 和 R 同等重要.

3.2 结果分析

实验结果如表 5 所示.实验数据分析如下.

(1)通过对比实验 3 和实验 4 的数据发现,双层条件随机场模型的识别结果明显高于单层条件随机场模型的识别结果.原因是在单层模型中,一些嵌套了赛事名简称、球队名简称和简单地名的复杂命名实体没有被识别出来,比如赛事名"2014 年法网女单决赛"、参赛球队名"浙江稠州银行男篮"、比赛场馆名"沈阳奥体中心五里河体育场".

(2)通过对比实验 1 和实验 4 的数据看出,与隐马尔科夫模型的识别结果相比,双层条件随机场模型识别的准确率、召回率和 F 值分别提高了 8%,10%,9%.原因是在隐马尔科夫模型中,只考虑融合了词和词性两种特征,而忽略了前缀、常用字以及左右边界词特征,这使得模型既不能较好的描述命名实体的内部结构也无法充分利用丰富的上下文信息,造成识别效果较差.

(3)通过对比实验 1 和实验 2 数据得知,隐马尔科夫模型和规则相结合方法的识别效果相对于单纯使用隐马尔科夫模型,其准确率、召回率和 F 值都有大幅提升.原因是在隐马尔科夫模型识别后,利用制定的领域规则对输出结果进行后处理,从而提高了复杂命名实体的识别效果.

(4)通过对比实验 2 和实验 4 的数据表明,相对于双层条件随机场模型,隐马尔科夫模型和规则相结合方法的识别效果,其准确率、召回率和 F 值分别下降了 2%,4%和 3%.原因是制定的领域规则不够完善,难以覆盖所有的语言现象.

(5)通过综合对比所有实验数据得出结论,基于双层条件随机场模型的方法取得了最优的实验结果.该方法取得成功的原因有:一是条件随机场模型可以方便的融合多种语言特征,并且克服了隐马尔科夫模型的不足;二是双层条件随机场模型将识别过程分为两步,首先在低层模型中识别出简单命名实体,然后在高层

模型中识别出嵌套了简单命名实体的复杂命名实体.

表 5　设计实验结果

实验	实体类别	准确率/%	召回率/%	F 值/%	实验	实体类别	准确率/%	召回率/%	F 值/%
	赛事名	87.61	83.94	85.74		赛事名	94.78	89.65	92.14
实验 1	参赛球队名	92.04	89.35	90.67	实验 3	参赛球队名	94.21	93.42	93.81
	比赛场馆名	89.56	87.25	88.39		比赛场馆名	95.13	92.95	93.67
	平均	89.74	86.85	88.27		平均	94.71	92.01	93.20
	赛事名	95.84	90.45	93.06		赛事名	97.83	96.37	97.09
实验 2	参赛球队名	95.27	94.86	95.06	实验 4	参赛球队名	98.12	97.50	97.81
	比赛场馆名	96.19	93.76	94.96		比赛场馆名	98.59	97.49	98.03
	平均	95.76	93.02	94.36		平均	98.18	97.12	97.64

4　结　论

本文针对复杂体育赛事命名实体的特点，提出了一种基于双层条件随机场模型的体育赛事命名实体识别方法. 该方法将命名实体识别过程分为两步，同时把多种语言学特征融合到条件随机场模型中，通过对真实体育赛事语料的开放测试表明，该方法取得比较理想的识别效果. 但在实验中发现了两个问题：一是漏识问题，比如赛事名“第 16 届 CUBA 中国大学生篮球联赛(东南赛区)冠亚军决赛”没有被识别出来；二是 CRF 对训练数据存在过拟合问题.

在下一步的工作中，一是要结合语义特征信息来提高命名实体识别效果以解决命名实体的漏识问题；二是采用新的 CRF 参数训练算法以消除对训练数据的过拟合问题；三是要积极开展体育赛事领域的深入研究，比如体育赛事搜索、体育赛事新闻推荐、体育赛事结果预测等.

参　考　文　献

[1] 蒋才智，王　浩. 基于知网的贝叶斯中文人名识别[J]. 南京大学学报：自然科学版，2012，48(2)：147-153.

[2] 李丽双，党延忠. CRF 与规则相结合的中文地名识别[J]. 大连理工大学学报，2012，52(2)：285-289.

[3] 胡万亭，杨　燕. 一种基于词频统计的组织机构名识别方法[J]. 计算机应用研究，2013，30(7)：2014-2016.

[4] 何炎祥. 基于 CRF 和规则相结合的地理命名实体识别方法[J]. 计算机应用与软件，2015，32(1)：179-185.

[5] 何琳娜，杨志豪. 基于特征耦合泛化的药名实体识别[J]. 中文信息学报，2014，28(2)：72-77.

[6] 梅　丰，孙承杰. 面向网络文本的中文产品命名实体识别[J]. 郑州大学学报：理学版，2010，42(1)：62-66.

[7] 杨永贵. 中文信息抽取关键技术研究与实现[D]. 北京：北京邮电大学，2008.

[8] 高国洋. 体育领域信息抽取系统的研究[D]. 北京：华北电力大学，2009.

[9] 汪　泱. 基于条件随机场的哈萨克语基本短语自动识别[J]. 计算机工程与设计，2014，35(10)：3602-3607.

[10] 谷　川，周宏宇. 融合多特征的中文产品命名实体识别[J]. 科学技术与工程，2013，13(31)：9417-9421.

[11] 于江德，葛彦强. 基于条件随机场的汉语词性标注[J]. 微电子学与计算机，2011，28(10)：63-66.

[12] 郭剑毅，薛征山. 基于层叠条件随机场的旅游领域命名实体识别[J]. 中文信息学报，2009，23(5)：47-52.

Research on Named Entity Recognition in Sports Events Filed

GU Chuan[a], SONG Xu[b]

(a. School of Software Engineering; b. School of Computer and Information Engineering, Anyang Normal University, Anyang 455000, China)

Abstract: In order to accurately recognize the complex sports events named entities in Chinese text, this paper presents a method of named entity recognition based on cascaded conditional random fields. In the proposed method, simple named entities are firstly recognized by lower model and then complex named entities nesting simple sports events named entity such as event name, team name and venue name are recognized by higher model. In open test on large-scale corpus, its F-measure of event name, team name and venue name is 97.09%, 97.81% and 98.03%.

Keywords: named entity recognition; sports events filed; cascaded conditional random fields