基于层次特征的藏文人名识别研究*

刘飞飞[1,2]，王志娟[1,2†]
(1. 中央民族大学 信息工程学院，北京 100081；2. 国家语言资源监测与研究中心 少数民族语言分中心，北京 100081)

摘 要：为了提高藏文人名识别的效果，提出了结合三层的层次特征的藏文人名识别算法。提出了无须分词，仅在藏文音节粒度上，基于藏文人名三层特征：内部特征、上下文信息、并列关系特征，利用条件随机场(conditional random fields，CRF)算法进行藏文人名识别的研究。首先将人名的内部和上下文特征作为 CRF 特征，然后将人名并列关系特征设计为规则进一步提高识别效果。在不影响准确率的情况下，最终将人名识别的召回率提高了 10.43%，综合 *F* 值达到了 95.02%。其中对于藏族人名的 *F* 值提升了 11%，音译人名识别的 *F* 值达到了 94.09%。实验结果表明，该方法可以有效提升藏文人名的识别效果。

关键词：人名识别；层次特征；藏文；条件随机场
中图分类号：TP391.1 **文献标志码**：A **文章编号**：1001-3695(2018)09-2583-05
doi:10.3969/j.issn.1001-3695.2018.09.005

Research on recognition of Tibetan names based on hierarchical features

Liu Feifei[1,2], Wang Zhijuan[1,2†]
(1. *School of Information Engineering, Minzu University of China, Beijing* 100081, *China*; 2. *National Language Resource Monitoring & Research Center of Minority Languages, Beijing* 100081, *China*)

Abstract: In order to improve the effect of Tibetan name recognition, this paper designed the algorithm based on three levels of hierarchical features. It proposed a three-layer feature, which was based on the Tibetan character name without word- segmentation. The three-layer feature included internal features, the context information and the parallel relations feature. It used the conditional random fields (CRF) algorithm to identify the Tibetan name research. First, it considered the internal and context characteristics of the name as a CRF feature, and then considered the relationship between names as the rule to further improved the recognition effect. The recall was increased 10.43% and the *F*-value will reach 95.02%. Experiment shows that the method achieves a very good effect for recognition of Tibetan names.

Key words: recognition of names; hierarchical features; Tibetan; conditional random field

命名实体(named entity)出现于第六次信息理解会议(MUC-6)(Grishman & Sundheim,1996)。命名实体识别作为信息抽取的一项子任务，意在识别文本中具有特定意义的实体，包括人名、地名、组织机构名，以及数字表达式，包括时间、日期、金钱和百分比表达式等[1]。

在藏文自然语言处理领域，人名识别效果直接影响到信息抽取方面的效果。当前对于藏文人名识别方法可以分为基于规则、规则与统计相结合、基于统计机器学习三种。基于规则的藏文人名识别方法是对该领域研究之初采用的主要方法，是根据格助词特征、人名词典特征、人名边界特征、人名高频词特征等设置规则，以实现对藏文人名的识别[2,3]。规则与统计相结合的藏文人名识别除了对于规则的利用之外，还结合了一定的统计方法，如互信息，分析人名与上下文的关联性[4]，以实现藏文人名的识别。在利用藏文人名规则的基础上，还利用近几年对于藏文人名识别的研究主要集中在统计学习方法方面，用到的统计学习方法包括条件随机场[5,6]、最大熵[6]、感知机[7]模型，对于藏文人名识别准确率能够在 90% 以上。然而对于藏文人名的识别还是存在一些问题，如音译人名识别的问题、较复杂的涉及宗教领域的藏文人名识别困难。另外，目前的藏文人名识别用到的特征范围局限在人名的上下文层次，对于人名的位置关系并未利用。

本文针对上述藏文人名的识别问题，设计了基于条件随机场统计模型，利用三层特征信息，即藏文人名的内部音节特征、上下文边界特征、人名的并列关系特征的藏文人名识别方案，以提高藏文人名识别的效果。

1 命名实体识别研究

1.1 命名实体识别方法汇总

命名实体识别的方法可以分为基于规则的方法和基于统计机器学习的方法两大类。

基于规则和词典的方法[8,9]是命名实体识别中最早使用的方法。一般而言，当提取的规则能够比较精确地反映语言现象时，基于规则的方法性能要优于基于统计的方法。但是这些规则往往依赖于具体语言、领域和文本风格，编制过程耗时长且难以涵盖所有的语言现象，特别容易产生错误，系统可移植性不好，对于不同的系统需要语言学专家重新书写规则。基于规则的方法另外一个缺点是代价太大，存在系统建设周期长、移植性差，而且需要建立不同领域知识库作为辅助以提高系统识别能力等问题。

基于统计机器学习的命名实体识别方法包括监督学习、半监督学习、无监督学习三类。监督学习(supervised learning,

收稿日期：2017-05-05；**修回日期**：2017-06-27 **基金项目**：国家自然科学基金重点资助项目(61331013)；国家语委科研项目(WT125-46)；中央民族大学一流大学一流学科研究生自主科研项目(10301-0170040601-184)

作者简介：刘飞飞(1993-)，女，硕士，主要研究方向为自然语言处理；王志娟(1977-)，女(通信作者)，副教授，博士，主要研究方向为自然语言处理(wangzj_muc@126.com).

SL)技术包括隐马尔可夫模型(HMM)[10,11]、最大熵模型(ME)[12]、支持向量机(SVM)[13]以及条件随机场(CRF)[14~17]。通常包括读取大的带注释的语料库,记住实体列表并基于区分特征创建消歧规则的系统。主要做法是通过对训练语料所包含的语言信息进行统计和分析,从训练语料中挖掘出特征。半监督式学习是综合利用有类标的数据和没有类标的数据来生成合适的分类函数,主要技术称做 bootstrapping。它也包含了一些监督式学习的方法,例如都需要从一系列种子来开始学习的过程,系统再用新找到的实体作为新的种子,重复地在文本中进行搜索的过程并寻找新的实例。与监督方法的性能相比,半监督的命名实体识别[18,19]性能具有很大的竞争力。无监督式学习[20]是直接对输入数据集进行建模。最典型的方法就是聚类,基本上都依赖于词汇资源(如词汇网络)、词汇特征和从大量未标注的语料库中计算而得到的统计概率。

另外,近几年随着神经网络算法的应用和发展,在命名实体识别领域也有了相关应用研究[16,21],且取得的研究结果表明基于神经网络算法的命名实体识别效果显著。

1.2 部分语种的命名实体识别现状

英文命名实体识别的研究开始较早,在1997年的MUC-7会议中,英文命名实体测试的 F 值就达到了93.39%[22],识别结果已经达到较高水平。中文的命名实体识别起步晚于英文,研究的开始阶段,基于条件随机场算法的命名实体识别[17]研究在MSRA评测集中 F 值可达到85.14%,近期研究[23,24]将递归神经网络与条件随机场相结合,在SIGHAN评测中 F 值达到90.51%。另外,越南语[25]、僧伽罗语[26]、葡萄牙语[27]、西班牙语[27]、印度语[27]、捷克语[28]等语料资源较少语种的命名实体识别的研究也在发展中。

鉴于不同语言的结构、用法等的差别,命名实体识别方法也需要针对语言现状进行分析。但是目前也有跨语言命名实体识别系统的研究,文献[29]实现了对五种语言的命名实体识别系统构建,包括英语、西班牙语、荷兰语、巴斯克语和德语。

藏文属汉藏语系藏缅语族藏语支,为藏族使用的主要语言。它是藏文分词、机器翻译、跨语言检索和文档摘要等自然语言处理中应用的关键技术。目前对于藏文命名实体识别的研究也在推进中,具体的研究现状将在第2章中详细介绍。

2 藏文人名识别研究现状

2.1 藏文人名识别方法

藏文命名实体识别包括对于藏文人名、地名、组织机构名等的识别。本文以藏文人名作为研究对象,并对目前藏文人名识别方法总结如下。为清楚表示目前藏文人名识别中的研究成果,根据已查的文献总结如表1所示。

表1 藏文人名识别研究汇总

方法类别	方法说明	使用语料	识别结果(F值)/%	需要分词
基于规则	Yu等人[2]:利用格助词规则,结合词典、训练策略等特征	2007年2月西藏日报 2008年12月中国西藏网	86.91	是
	Sun等人[3]:基于多特征,包括藏文人名词典匹配、边界特征、上下文特征、人名高频词等	2009年藏文网页文本(400句)	83.91	是
统计与规则相结合	窦嵘等人[4]:互信息统计方法,结合格助词规则和人名词典	训练集:2007年西藏日报(30.2 MB) 封闭测试:2007年2月西藏日报 开放测试:中国共产党新闻网藏文版(100篇)	封闭测试:93.55 开放测试:87.92	是
	加羊吉等人[6]:条件随机场结合最大熵模型	训练集:《西藏日报》2007年1月的语料(3.5 MB) 测试集:《西藏日报》2007年2月1日至20日的语料(2.1 MB)	93.08 (CRF:92.38 ME:91.54)	是
基于机器学习	华却才让等人[7]:感知机模型	训练集:网络藏文文本(15 000条句子,包含9 655个人名) 测试集:网络藏文文本(1 016条句子,包含853个人名)	88.04	否
	康才畯等[5]:条件随机场	训练集:社科院人工分词平衡语料(4万余字,共1 951个藏文人名) 封闭测试:训练集的10% 开放测试1:中学藏文教材 开放测试2:藏译本《水浒传》节选	封闭测试:94.31 开放测试1:91.26 开放测试2:84.70	否

藏文人名识别作为藏文命名实体研究的一部分,主要出现在对藏文命名实体识别的研究论述中。最初是金明等人[30]提出的融合规则方法和基于HMM(hidden Markov model)的统计机器学习方法,对藏文命名实体识别提出解决思路。

早期的藏文命名实体识别的研究主要采用的是基于规则的方法。Yu等人[2]提出的利用格助词语法的藏文命名实体识别方法,在大量语料中提取藏文人名的边界特征,基于藏文格助词规则,结合词典、训练策略等特征,在网站文本测试中 F 值达到86.91%。Sun等人[3]提出的基于多特征的藏族人名抽取方法,结合藏文人名词典匹配、边界特征、上下文特征、人名高频词等多个特征依次判别,实现藏文人名的抽取,实验测试 F 值达到83.91%。窦嵘等人[4]采用统计与规则相结合的方法,运用互信息的统计方法,结合格助词规则与人名词典解决藏文人名识别,测试中 F 值最高可以达到93.55%。

近几年对于藏文命名实体识别的研究以基于监督学习为主,包括条件随机场、最大熵模型与感知机模型等。加羊吉等人[6]结合条件随机场和最大熵模型两种机器学习方法,在西藏日报的语料上进行实验,获得了93.08%的 F 值;华却才让等人[7]利用感知机模型,并研究了基于分词与基于音节两种特征下的实体识别结果,在藏文人名识别上达到了88.04%的 F 值;康才畯等人[5]以条件随机场作为基础识别器,利用音节字符在人名方面的不同作用划分特征集合,在封闭测试中 F 值达到了94.31%,开放测试中 F 值也达到了91.26%。

由于目前没有针对藏文命名实体识别的评测工具,所以不同研究中使用语料、语料识别结果均有差异。虽然各类研究中使用的语料和方法均不相同,仅根据 F 值高低判断方法好坏有些片面,但是康才畯等人在不用分词的情况下达到了94.31%的 F 值,一定程度上说明了条件随机场这一方法的有效性。本文也将采用条件随机场作为基本进行研究,而且无须分词,在藏文音节层面进行研究,在减少分词操作的同时避免了分词错误引发的识别错误。

2.2 藏文人名识别使用的特征

鉴于不同研究对于特征的表述方式存在差别,有利用相同特征但是表述不同的情况,本文对这些特征进行了整理分类,按照特征层次总结如表2所示。

根据表2对于藏文人名识别中用到的特征信息的总结,发现用到人名本身与人名上下文的特征已经较多,但是并未有涉及人名在句子中的位置特征。另外,对于人名本身的特征时,对

于藏文人名的特征，如藏文高频词，已有利用且取得效果明显，但是对于藏文人名中的译名特征并未有涉及，以及对于人名上下文特征的利用时，对普通的人名上下文进行了总结整理，如一般称号、职业、荣誉和亲属关系，但是针对藏文文本中的文化特性，存在一些不常用的藏文人名上下文词，如班禅、仁波切等。

表 2　藏文人名识别使用特征

特征层次	特征种类以及说明	特征应用实例
人名内部	人名词典特征	Yu 等人：3 530 个藏族常用人名；Sun 等人：《常用藏语人名地名词典》，以及另外的 3 119 个藏文人名；窦嵘等人：3 727 个藏族人名；加羊吉等人：2 058 个藏文人名
	人名用字特征	Sun 等人：102 个藏文人名高频词；加羊吉等人：收录 273 个译名用字
人名上下文	格助词特征	Yu、窦嵘等人
	边界特征	Sun 等人：110 个人名边界词条

针对上述情况，本文在对于藏文人名的识别特征的选择时，将汉族音译人名的特征加入到人名的内部特征部分；以藏族宗教领域的称谓词扩充藏文人名的上下文特征；将藏文句子中并列人名的特征作为位置规则，以期望进一步增强藏文人名的识别效果。

2.3　藏文人名识别存在的问题

现有的藏文人名识别已经取得了一定的成功，但仍存在如下问题：

a）音译人名不能正确识别。

如：བང་ཡན་/PER　ལ་ཡུལ་　དྲ་སོག

王艳　　家乡　回到。

b）人名边界为特殊宗教名词未能识别。

如：པན་ཆེན་སྐུ་སྡེད་གསུམ་པ་ བློ་བཟང་བསྟན་འཛིན་ /PERནོ་གཅིག་པའི་ཚེས་བཅུ་དགུར་སྐྱེ་འཕྲུངས།

班禅 三世　　罗桑丹珠　　一月四日　　诞生。

c）句子中存在并列关系的人名不能全部识别。

如：བློན་པོ་མཚར་སྒོང་བཙན་ /PER ལ་དུ་ཐུ་ཡོད།མཚར་བཙན་སྐུ་རྩོམ་བྱ/PER ।

大臣 禄东赞 有五个儿子：噶尔·赞悉若、

མཚར་ཁྲི་འཛིང་ /PER ।མཚར་བཙན་རྩོད/PER ।བློན་ཆེ་མདོ་ཡེ/PER ।

噶尔·论钦陵、噶尔·赞婆、噶尔·悉多于、

བློན་འབབ་ལོན་ར/PER ।

噶尔·勃伦赞刃。

其中横线标注的五个儿子的人名并列显示，而识别结果为 བློན་པོ་མཚར་སྒོང་བཙན་ /PERལ་བྱ་ཐུ་ཡོད།མཚར་བཙན་སྐུ་རྩོམ་བྱ།མཚར་ཁྲི་འཛིང་ /PER ।མཚར་བཙན་རྩོད་དབྱིན་ཆེ་མདོ་ཡེ།བློན་འབབ་ལོན་ར། ，仅识别出一个人名。

面对这些问题，本文提出了相应的解决方案如下：

a）对于音译人名的识别问题，加入了汉族人名常用字即汉姓音译特征，借助汉姓人名的特征，实现对音译人名中汉姓人名的识别效果的提高。

b）藏族宗教领域人名识别问题，原因在于藏文宗教领域人名的上下文大多为特殊的称谓词，所以本文除了普通的人名上下文用词（普通称谓词，如大臣、主席、经理），对于藏文文本中常用的宗教文化领域的人名上下文进行了汇总整理，对上下文特征词进行了补充。

c）句子中存在并列关系的人名不能全部识别的问题，本文在实验分析中发现了存在句子中多个并列人名不能全部识别，对于此类问题，设计了人名位置规则，将进一步改进藏文人名识别效果。

3　基于层次特征藏文人名识别研究

3.1　藏文人名的层次特征

藏语属于汉藏语系，理论上，汉语中使用的自然语言处理方法都可以用在藏语信息处理中，但在实际使用过程中必须考虑藏语中存在的具体问题。本文利用分层的藏文人名特征实现对藏文人名的识别，主要分为人名内部特征、人名上下文信息、人名并列关系特征三层。

3.1.1　人名内部特征

在藏文人名识别中藏文人名包括藏族人名、藏文译名。不同语言，人名的内部特征不同。比如英文人名首字母大写，汉族人名的“姓氏 + 名字”组合结构。藏族人名也有自己独特的特征。一般没有严格意义上的姓氏，旧西藏贵族和宗教界认识名字中的家族名、寺庙名等类似汉族人名中的姓，大多数普通人的名字中没有“姓氏”的部分。另外，藏族人名的书写结构与普通藏文词并无区别。

根据以上对藏文人名特征的描述发现，藏族人名内部没有结构特征和字形特征可供参考。但是对比中文名中常用“伟、丽…”等常用汉字，藏文人名也存在常用藏文词。所以可通过高频藏文音节作为特征借以提高藏族人名的识别准确率。

根据前期收集的 10 460 个藏族人名，对其中的高频音节进行统计，并经过人工校对分析，最终选择了出现频率在 1% 以上的 97 个藏族人名高频音节。藏族人名高频音节示例如表 3 所示。

表 3　藏族人名高频音节示例

藏文人名常用词	翻译	出现频率/%
རྫེ་དཔོ་	才让	1.89
བཀྲ་ཤིས་	扎西	1.89
བསྟན་འཛིན་	丹增	1.75
རྡོ་རྗེ་	道吉	1.75
སྒྲོལ་མ་	卓玛	1.73

藏文译名中很大一部分是汉译名。藏文音译的中文人名，最常用的音节就是姓氏，而且译名用字有一定规律，大部分都是藏语中不常用的字，所以可将音译的汉姓作为识别藏文音译的中文人名的特征。汉姓选择百家姓，其中单姓 444 个，复姓 60 个，共 504 个[31]，将其中的汉姓字音译为藏文音节，因为汉姓中的同音情况，音译获得包含 291 个汉姓音译的汉姓音译表。部分汉姓音译如表 4 所示。

表 4　部分汉姓音译

音译藏文	对应汉姓
འང་	王，汪
ལི་	李，里，黎，栗，厉
ཛང་	张，章
ཡན་	严，颜，燕，晏，阎，言
ཨུ་	吴，伍，巫，乌，武
གུང་	巩，宫，弓，龚，贡，公

3.1.2　人名上下文特征

在文献[2]中对人名的边界信息归纳统计，并认为人名的左边界一般是职务、职业、称谓等词，人名的右边界一般是格助词，也有敬语、语气助词等情况。笔者在语料中人名的边界信息进行统计时也发现了这种情况。除此之外，在语料中发现人名的上下文还会出现与藏文宗教文化领域相关的词，如ཁྲི་འདོན་（坐床）、རྫོང་འདས་（圆寂）等。所以，将人名的上下文特征总结为格助词特征、普通人名上下文特征（包括职务、一般称谓等）、藏文文化领域特殊上下文特征（如班禅、坐床等）三类。对于这三类上下文特征的利用，鉴于格助词特征一般紧邻人名且用法固定，可以通过条件随机场的训练集合识别窗口加以利用，而对于普通人名上下文特征和藏文文化领域上下文特征，需要通过预先的准备，获取人名上下文常用词表。

本文共收集到包括 34 个藏文文化领域上下文特征词的共

132 个人名上下文特征词表。宗教领域的上下文特征词如表 5 所示。

表 5 宗教领域特征词示例

特征词	翻译	含义
ཁྲི་འདྲོན་	坐床	藏传佛教中喇嘛活佛“转世”继位的仪式
ཐུགས་འདུས་	圆寂	佛界语,指僧人死后升天
བཙན་ཞེན་	班禅	藏传佛教信徒一般认为班禅是“月巴墨佛”即阿弥陀佛的化身
རིན་པོ་ཆེ	仁波切	藏文(rin-po-che)的音译,意指“珍宝”或“宝贝”

3.1.3 人名并列关系特征

在分析人名识别结果时发现,在语料中并列关系人名仅能识别部分人名的情况。这种情况的出现原因为可识别的部分人名识别特征明显,而另外的人名缺少特征无法识别。这种情况在中文识别中可以通过顿号,将与人名有并列关系的实体正确识别,而藏文标点类型极少,一般用单垂线“|”表示顿号、句号、逗号等情况,所以,对于藏文人名的并列情况,需要先判断并列关系。

本文对于人名并列关系的判断以句子的音节数为基础。首先对包括 18 789 个句子的 5.3 MB 藏文文本进行统计,发现句子的平均音节数是 21.6 个,其中句子音节数不超过 5 的占总数的 10.8%。在对此类句子进行抽样分析时发现此类短句绝大部分为短句、词组等,并非完整的藏文句。所以本文以句子音节数为 5 作为分界点,不超过 5 个藏文音节的判定为藏文短句。出现 2 个以上的藏文短句则确定此处存在并列关系。

确定并列关系存在后,根据如下规则进行人名的识别:如果存在藏文短句,其上下文也是藏文短句,三者中存在标记为人名的情况,则表示将此藏文短句为人名。本文将利用这种并列关系特征以改进藏文人名识别的效果。研究之后根据人名并列关系特征规则判定是否为人名。

3.1.4 人名层次特征总结

将本文用到的人名层次特征总结如表 6 所示。

表 6 人名层次特征

特征类型		特征规模
人名内部特征	藏文人名高频音节	97 个藏族人名高频音节
	汉姓音译音节	291 个汉姓音译藏文
人名上下文特征	上下文称谓特征	132 个人名上下文特征词表(包括 34 个藏文文化领域上下文特征词)
人名并列特征	并列关系特征	识别规则

3.2 基于层次特征的藏文人名识别

条件随机场模型(conditional random field,CRF)是 Lafferty 等人[32]于 2001 年在最大熵模型和隐马尔可夫模型的基础上提出的一种无向图学习模型,是一种用于标记和切分有序数据的条件概率模型。它没有独立性假设,可任意选择特征,并且对所有特征进行全局归一化,从而得到全局最优解。

本文在藏文人名的识别研究中,以条件随机场作为基本的识别器,将人名内部特征、人名上下文特征作为信息,并设计特征模板,以实现对藏文人名的识别。其具体说明如表 7 所示。其中%x[row,col]表示确定输入数据中的一个 token,row 表示确定与当前的 token 的相对行数,col 用于确定绝对列数。

表 7 特征模板设计

层次特征	特征描述	特征模板部分示例
人名内部特征	音节与是否为藏文人名高频音节、是否为音译汉姓的组合特征	%x[0,0]/%x[0,1]:当前音节与 HFT 的组合特征 %x[0,0]/%x[0,2]:当前音节与 HFC 的组合特征
人名上下文特征	音节与之前的音节之间是否满足人名与人名上下文关系特征	%x[0,0]/%x[-1,3]:当前音节与前一音节的特征 %x[0,0]/%x[-1,3]/%x[-2,3]:当前音节与前两个音节的特征

而对于人名的位置特征,本文设计为对基于条件随机场的识别标注结果,利用人名并列关系规则进行二次标注修正,以实现对于藏文人名的识别效果的提升。

4 实验与分析

4.1 实验语料与评测指标说明

本实验的语料是选择人民网藏文版的网页文本,文本类型涉及新闻、政治、宗教、文化各个领域,训练集的大小为 3.78 MB,共包含 2 557 个人名,测试集的大小为 1.51 MB,包含 908 个人名。藏族人名和译名的数量表示如表 8 所示。

表 8 语料人名数目统计

语料集合	语料集合大小/MB	藏族人名数目	音译人名数目
训练语料	3.78	975	1 582
测试语料	1.51	388	520

测试中本文采取了三个评测指标:

准确率 = 正确识别的人名总数/识别出的人名总数

召回率 = 正确识别的人名总数/测试集中存在的人名总数

F 值 = 正确率 × 召回率 ×2/(正确率 + 召回率)

4.2 各类特征对藏文人名识别的影响

在实验中,首先测试仅标注音节信息的情况下 CRF 的藏文人名识别性能,并以此作为基准;之后依次加入各类特征,以分析各类特征对藏文人名识别是否有效,以及不同特征对于藏族人名与音译人名识别效果的影响。鉴于位置关系的特征用于二次标注的修正过程,在此处本文不单独对其进行分析,其影响性将在对层次特征的分析中体现。

各类特征对于藏文人名识别的实验结果如表 9 所示。本文将藏族人名与音译人名分开讨论。

表 9 各类特征对藏文人名识别的影响(藏族人名/音译人名)

增加特征	准确率/%	召回率/%	F 值/%
基线	94.39/96.83	78.09/82.12	85.47/88.87
内部特征(汉姓音译)	94.58/96.53	80.93/85.58	87.22/90.72
内部特征(藏族人名高频音节)	95.24/96.90	82.47/84.23	88.40/90.12
上下文特征	95.52/96.73	82.47/85.38	88.52/90.70

a)汉姓音译特征的影响。在增加汉姓音译特征之后,藏族人名和音译人名召回率和 F 值均有所提升。藏文人名的召回率提升了 2.84%,而音译人名召回率提升了 3.46%,这说明藏文汉译人名的姓氏音译特征对于音译人名的识别效果更好。

b)藏族人名高频音节特征的影响。藏族人名高频音节特征的增加将藏族人名 F 值提升了 2.93%,音译人名提升了 1.25%,这体现了藏族人名高频音节特征可以更好地提升藏族人名的识别效果。

c)上下文特征的影响。增加上下文特征之后,藏族人名识别 F 值提升了 3.05%,音译人名的 F 值达到了 90.70%,提升了 1.83%。上下文特征对于两类藏文人名都有提升作用,尤其是上下文特征中增加的藏文文化领域称谓词对藏族人名的识别作用更明显。

根据结果显示,增加的每一种特征,对藏文人名识别效果均有所提升。

4.3 层次特征对藏文人名识别的影响

在明确了研究中采用的各类特征对藏文人名识别的积极效果后,接下来按照层次增加特征进行藏文人名识别结果。实验中依次叠加各层特征,以了解层次特征对于藏文人名识别的影响情况。实验结果如表 10 所示。

表 10 叠加各层次特征的结果(藏族人名/音译人名)

增加的特征层次	准确率/%	召回率/%	F 值/%
基线	94.39/96.83	78.09/82.12	85.47/88.87
内部特征	96.08/96.37	82.22/86.73	88.61/91.30
内部特征 + 上下文特征	95.97/98.32	89.18/89.81	94.15/93.87
内部特征 + 上下文特征 + 并列位置特征	99.45/98.12	92.78/90.38	96.00/94.09

根据表 10 的结果,分析各层特征对于藏文人名识别的影响。

a)在原实验基础上增加藏文人名的内部特征,召回率提升明显,藏族人名和音译人名均提升了超过 4%,综合 F 值提升了 3% 左右,这说明仅增加内部特征可以提升藏文人名识别的效果。

b)叠加内部特征与上下文特征之后,人名识别的准确率没有损失,且藏族人名和音译人名的召回率提高到了 89% 以上,F 值达到了 94% 左右,这说明结合内部与上下文层次的特征,在保证准确率的同时召回率有明显提高。

c)在人名内部特征和上下文特征之后,利用位置特征规则进行再次识别。藏族人名 F 值提升明显,上升至 96%,音译人名也有细微提升幅度。

在条件随机场识别模型基础上,叠加了人名内部特征、上下文特征与并列位置特征,人名识别的效果提升明显,体现了层次特征对藏文人名识别的积极影响。

4.4　基于条件随机场的藏文人名识别研究的对比分析

目前已经有基于条件随机场的藏文人名识别方法,本文是基于此模型,但是特征与语料存在差异。表 11 对几种方法进行了对比分析。鉴于其他的研究并未区分藏文人名中的藏族人名和音译人名,对比藏文人名的 F 值。通过比较 F 值可以看出,本文使用的方法优于已有的同类研究。

表 11　基于条件随机场方法的藏文人名识别研究对比

方法	训练集	测试集	F 值/%
加羊吉等人:CRF + ME	3.5 MB	2.1 MB	93.08
康才畯等人:CRF + 多标签集	四万余字	四千字左右	94.31
本文方法:CRF + 层次特征	3.78 MB	1.51 MB	95.02

5　结束语

本文从分析藏文人名的特点出发,从藏文人名内部特征、人名上下文、人名并列关系这三层对藏文人名的特点进行了分析,提出了融合三层层次特征的藏文人名识别算法,在无须分词的情况下,以藏文音节作为基本粒度,利用条件随机场模型,实现了藏文人名的识别。在人名内部特征方面,除了藏族人名使用的高频音节,还将汉姓音译作为特征加入,改进了音译汉族人名的识别;在人名上下文层面,将人名上文下文出现的特征词进行了整理汇总,并将其作为特征加入,丰富藏文人名的上下文特征;另外,设计人名并列关系特征作为人名的识别规则,将并列关系的人名进行了二次标注处理。根据实验结果发现,三层特征相结合对于藏文人名的识别有改进效果。

下一步将外国音译人名的特征进行总结,以提高外国音译人名的识别效果。另外,建立藏文人名词典信息,完善人名上下文特征词表,并加入句子的语义信息,以期实现更好的藏文人名识别效果。

参考文献:

[1] Nadeau D,Sekine S. A survey of named entity recognition and classification[J]. Journal of Linguisticae Investigationes,2007,30(1):3-26.

[2] Yu Hongzhi,Jiang Tao,Ma Ning. Named entity recognition for Tibetan texts using case-auxiliary grammars[C]//Lecture Notes in Engineering & Computer Science. 2010.

[3] Sun Yuan,Yan Xiaodong,Zhao Xiaobing,*et al*. Research on automatic recognition of Tibetan personal names based on multi-features[C]//Proc of the 6th International Conference on Natural Language Processing and Knowledge Engineering. Piscataway,NJ:IEEE Press,2010:1-5.

[4] 窦嵘,加羊吉,黄伟. 统计与规则相结合的藏文人名自动识别研究[J]. 长春工程学院学报:自然科学版,2010,11(2):113-115.

[5] 康才畯,龙从军,江荻. 基于条件随机场的藏文人名识别研究[J]. 计算机工程与应用,2015,51(3):109-111.

[6] 加羊吉,李亚超,宗成庆,等. 最大熵和条件随机场模型相融合的藏文人名识别[J]. 中文信息学报,2014,28(1):107-112.

[7] 华却才让,姜文斌,赵海兴,等. 基于感知机模型藏文命名实体识别[J]. 计算机工程与应用,2014,50(15):172-176.

[8] 周昆. 基于规则的命名实体识别研究[D]. 合肥:合肥工业大学,2010.

[9] Alfred R,Leong L C,On C K,*et al*. Malay named entity recognition based on rule-based approach[J]. International Journal of Machine Learning & Computing,2014,4(3):300-306.

[10] Biswas S,Mohanty S,Mishra S P. A hybrid oriya named entity recognition system:integrating HMM with MaxEnt[C]//Proc of the 2nd International Conference on Emerging Trends in Engineering & Technology. Washington DC:IEEE Computer Society,2009:639-643.

[11] 俞鸿魁,张华平,刘群,等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报,2006,27(2):87-94.

[12] Chieu H L,Ng H T. Named entity recognition with a maximum entropy approach[C]//Proc of the 7th Conference on Natural Language Learning at HLT-NAACL. Stroudsburg,PA:Association for Computational Linguistics,2003:160-163.

[13] Benajiba Y,Diab M,Rosso P. Arabic named entity recognition:an SVM-based approach[C]//Proc of International Arab Conference on Information Technology. 2009.

[14] McCallum A,Li Wei. Early results for named entity recognition with conditional random fields,feature induction and Web-enhanced lexicons[C]//Proc of the 7th Conference on Natural Language Learning at HLT-NAACL. Stroudsburg,PA:Association for Computational Linguistics,2003:188-191.

[15] Zhang Yuejie,Xu Zhiting,Zhang Tao. Fusion of multiple features for Chinese named entity recognition based on CRF model[C]//Proc of the 4th Asia Information Retrieval Conference on Information Retrieval Technology. Berlin:Springer-Verlag,2008:95-106.

[16] Chiu J P C,Nichols E. Named entity recognition with bidirectional LSTM-CNNs[EB/OL]. (2016-07-19). https://arxiv.org/abs/1511.08308v5.

[17] Chen Wenliang,Zhang Yujie,Isahara H. Chinese named entity recognition with conditional random fields[Z]. 2006:118-121.

[18] Liao Wenhui,Veeramachaneni S. A simple semi-supervised algorithm for named entity recognition[C]//Proc of the NAACLHLT Workshop on Semi-Supervised Learning for Natural Language Processing. Stroudsburg,PA:Association for Computational Linguistics,2009:58-65.

[19] Nadeau D. Semi-supervised named entity recognition:learning to recognize 100 entity types with little supervision[D]. Ontario:University of Ottawa,2007.

[20] Shinyama Y,Sekine S. Named entity discovery using comparable news articles[C]//Proc of International Conference on Computational Linguistics. Stroudsburg,PA:Association for Computational Linguistics,2004:848.

[21] Lample G,Ballesteros M,Subramanian S,*et al*. Neural architectures for named entity recognition[EB/OL]. (2016-04-07). https://arxiv:org/abs/1603.01360.

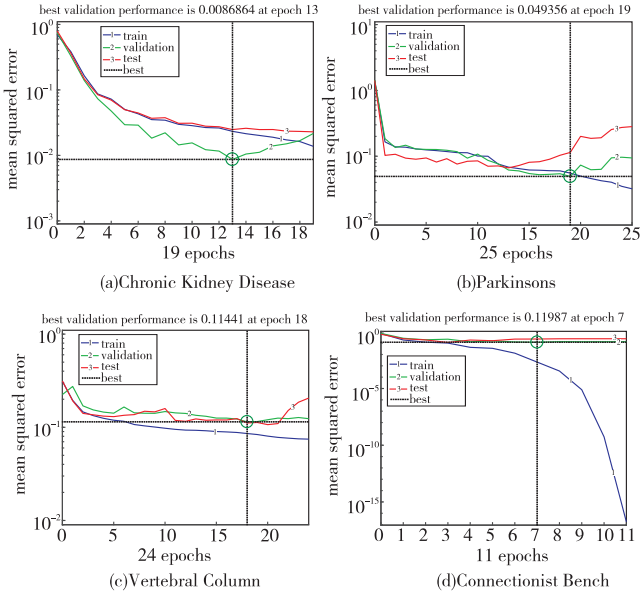
[22] Marsh E,Perzanowski D. MUC-7 evaluation of IE technology:overview of results[C]//Proc of the 7th Message Understanding Conference. 1998:20.

[23] Dong Chuanhai,Zhang Jiajun,Zong Chengqing,*et al*. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition[M]//Natural Language Understanding and Intelligent Applications. Berlin:Springer International Publishing,2016:239-250.

[24] Santos C N D,Guimarães V. Boosting named entity recognition with neural character embeddings[EB/OL]. (2015-07-31). https://arxiv.org/abs/1505.05008.

(下转第 2596 页)

模型的准确率、精度和召回率，降低神经网络模型的均值误差。RPN 算法运行时间的实验结果分析表明，RPN 算法比传统算法的运行时间最高降低了 62.27%，说明本文算法可以减少生成模型的时间。另外，本文提出的 RPN 算法存在实验流程较为复杂的问题，这将是下一步的工作和研究的方向。



(a)Chronic Kidney Disease (b)Parkinsons

(c)Vertebral Column (d)Connectionist Bench

图5 RPN算法MSE迭代过程

参考文献：

[1] 张超群，郑建国，王翔. 蜂群算法研究综述[J]. 计算机应用研究，2011，28(9)：3201-3205，3214.

[2] Gutierrez-Osuna R，Nagle H T. A method for evaluating data-preprocessing techniques for odour classification with an array of gas sensors [J]. IEEE Trans on Systems，Man，and Cybernetics，Part B：Cybernetics，1999，29(5)：626-632.

[3] Polikar R. Ensemble based systems in decision making[J]. IEEE Circuits and Systems Magazine，2006，6(3)：21-45.

[4] Rokach L. Ensemble-based classifiers[J]. Artificial Intelligence Review，2010，33(1-2)：1-39.

[5] Wang Xizhao，Xing Hangjie，Li Yan，*et al*. A study on relationship between generalization abilities and fuzziness of base classifiers in ensemble learning[J]. IEEE Trans on Fuzzy Systems，2015，23(5)：1638-1654.

[6] Freund Y，Schapire R E. Experiments with a new boosting algorithm [C]//Proc of the 13th International Conference on Machine Learning. 1996.

[7] Freund Y，Schapire R E. A desicion-theoretic generalization of on-line learning and an application to boosting[C]//Proc of the 2nd Annual European Conference on Computational Learning Theory. Orlando，FL：Academic Press，Inc，1995.

[8] Huang Chang，Wu Bo，Haizhou A I，*et al*. Omni-directional face detection based on real adaboost[C]//Proc of International Conference on Image Processing. Piscataway，NJ：IEEE Press，2004.

[9] Wu Shuqiong，Nagahashi H. Parameterized AdaBoost：introducing a parameter to speed up the training of real AdaBoost[J]. IEEE Signal Processing Letters，2014，21(6)：687-691.

[10] Schapire R E，Singer Y. Improved boosting algorithms using confidence-rated predictions[J]. Machine Learning，1999，37(3)：297-336.

[11] Friedman J，Hastie T，Tibshirani R. Additive logistic regression：a statistical view of boosting (with discussion and a rejoinder by the authors)[J]. The Annals of Statistics，2000，28(2)：337-407.

[12] Chawla N V，Lazarevic A，Hall L O，*et al*. SMOTEBoost：improving prediction of the minority class in boosting[C]//Proc of European Conference on Principles of Data Mining and Discovery in Databases. 2003：107-119.

[13] Seiffert C，Khoshgoftaar T M，Van Hulse J，*et al*. RUSBoost：a hybrid approach to alleviating class imbalance[J]. IEEE Trans on Systems，Man，and Cybernetics，Part A：Systems and Humans，2010，40(1)：185-197.

[14] Benesty J，Chen Jingdong，Huang Yiteng，*et al*. Pearson correlation coefficient[M]. Berlin：Springer，2009：1-4.

[15] Zhong Kai，Yang Qiqi，Zhu Song. New algebraic conditions for ISS of memristive neural networks with variable delays[J]. Neural Computing & Applications，2017，28(8)：2089-2097.

[16] 王小川. MATLAB 神经网络 43 个案例分析[M]. 北京：北京航空航天大学出版社，2013.

[17] Chen Yishi，Zhao Xing，Jia Xiuping. Spectral-spatial classification of hyperspectral data based on deep belief network[J]. IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing，2015，8(6)：2381-2392.

[18] Chen Yushi，Jiang Hanlu，Li Chunyang，*et al*. Deep feature extraction and classification of hyperspectral images based on convolutional neural networks[J]. IEEE Trans on Geoscience & Remote Sensing，2016，54(10)：6232-6251.

[19] Krizhevsky A，Sutskever I，Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Proc of the 25th International Conference on Neural Information Processing Systems. 2012：1097-1105.

[20] Jia P，Zhang M，Yu W，*et al*. Hyperspectral image feature extraction method based on sparse constraint convolutional neural network[Z]. 2017.

[21] Dr Soundarapandian. Chronic_kidney_disease[EB/OL]. [2017-03-06]. https://archive. ics. uci. edu/ml/datasets/Chronic_Kidney_Disease.

[22] Denver C. Parkinsons[EB/OL]. [2017-03]. https://archive. ics. uci. edu/ml/datasets/Parkinsons.

[23] Mota H D. Vertebral + Column[EB/OL]. [2017-03]. https://archive. ics. uci. edu/ml/datasets/Vertebral + Column.

[24] Venkat V，Chan K. A neural network methodology for process fault diagnosis[J]. AiChE Journal，2010，35(12)：1993-2002.

(上接第 2587 页)

[25] Pham Q H，Nguyen M L，Nguyen B T，*et al*. Semi-supervised learning for vietnamese named entity recognition using online conditional random fields[C]//Proc of the 7th ACL-IJCNLP Named Entities Workshop. 2015.

[26] Manamini S A P M，Ahamed A F，Rajapakshe R A E C，*et al*. Ananya：a named-entity-recognition (NER) system for Sinhala language [C]//Proc of Moratuwa Engineering Research Conference. Piscataway，NJ：IEEE Press，2016：30-35.

[27] Nongmeikapam K，Shangkhunem T，Chanu N M，*et al*. CRF based name entity recognition (NER) in Manipuri：a highly agglutinative Indian language[C]//Proc of the 2nd National Conference on Emerging Trends and Applications in Computer Science. Piscataway，NJ：IEEE Press，2011：1-6.

[28] Konkol M，Konopík M. CRF-based czech named entity recognizer and consolidation of czech NER research[C]//Proc of International Conference on Text，Speech，and Dialogue. Berlin：Springer，2013：153-160.

[29] Agerri R，Rigau G. Robust multilingual named entity recognition with shallow semi-supervised features[J]. Artificial Intelligence，2016，238(9)：63-82.

[30] 金明，杨欢欢，单广荣. 藏语命名实体识别研究[J]. 西北民族大学学报：自然科学版，2010，31(3)：49-52.

[31] 李捷译注. 百家姓[M]. 呼和浩特：远方出版社，2007.

[32] Lafferty J D，McCallum A，Pereira F，*et al*. Conditional random fields：probabilistic models for segmenting and labeling sequence data[C]//Proc of the 18th International Conference on Machine Learning. 2001：282-289.