基于条件随机场的藏文人名识别研究*

康才畯[1]，龙从军[2]，江荻[1,2]

（1. 上海师范大学 人文与传播学院，上海市 200234；2. 中国社科院 民族研究所，北京 100081）

摘要：本文基于条件随机场模型在字粒度上识别并切分藏文人名,其优势是可以较好地利用藏文人名在文本中出现的基本特征和上下文特征来确定藏文人名在文本序列中的边界。本文首先根据藏文人名自身的特点设定特征标签集，再利用条件随机场模型作为标注建模工具来进行训练和测试。从实验结果来看，该方法有较高的识别正确率，具有进一步研究的价值。下一步的改进需要扩充训练语料，并针对人名与一般词语同形现象进行特征标签集的优化。

关键词：藏文人名；条件随机场；特征标签集

中图分类号：TP391　　**文献标识码：**A

Tibetan Names Recognition Research based on CRF

Caijun Kang[1], Congjun Long[2], Di Jiang[1, 2]

(1. Humanities and Communications College, Shanghai Normal University, Shanghai 200234, China; 2. Institute of Ethnology & Anthropology, Chinese Academy of Social Sciences, Beijing 100081, China)

Abstract: The best feature of segmentation of Tibetan names based on Conditional Random Field (CRF) on the character level is making use of the the basic and context features of the Tibetan names. This article defines a feature tag set to fit in with the characters of Tibetan names, uses CRF as tagging model to train and test corpus data. The experimental result shows that the method has a high recognition rate and deserves further study. The next steps are to expand the corpus and optimize the tag set for the isomorphic phenomena of Tibetan names and general words.

Key words: Tibetan name; CRF; tag set

1 引言

分词技术是东亚语言自然语言处理的前提和基础，其中未登录词的识别是影响分词精度的一个重要方面，也是自动分词技术的难点之一。未登录词通常包括人名、地名、机构名等，往往与其前后的字词交叉组合，或与一般词汇同形，不仅增加了自身切分的难度，而且严重地干扰了相邻词的正确切分，从而大大地降低了分词的正确率。在藏文分词中，藏文人名在未登录词中占有很大比重[1]。因此，藏文人名的自动识别对于藏文未登录词识别以及藏文自动分词具有重要的意义。

针对人名的自动识别主要有三种：规则方法、统计方法以及规则与统计相结合的方法。基于规则的方法对人名的构成特征及上下文信息特征进行分析归纳，建立规则集。该方法具有很高的准确率，但召回率不高，在缺乏大规模语料库的时候，规则似乎是唯一可行的方法[2]。统计方法主要是建立统计模型对姓名语料库进行训练，得到候选字段作为姓名的概率，设定阈值从而判断是否为人名[3,4]。规则与统计相结合的方法，一方面通过概率计算减少规则方法的复杂性与盲目性，另一方面通过规则的复用，降低统计方法对语料库规模的要求[5]。目前的研究基本上都采取规则与统计相结合的方法，不同之处仅在于规则与统计的不

* **收稿日期：2013-07-27；定稿日期：2013-08-05**

基金项目：基于本体的多策略民汉机器翻译研究(No.61132009)

作者简介：康才畯(1980--)，男，博士研究生，主要研究方向为少数民族语言处理，计算语言学；龙从军（1978--），男，博士研究生，助理研究员，主要研究方向为现代藏语语法，计算语言学；江荻（1956--），男，博士，高级研究员，主要研究方向为汉藏语言学，现代藏语语法和计算语言学。

同侧重。

本文提出一种基于条件随机场的藏文人名识别方法。该方法围绕人名及其上下文设计特征标签集，然后通过条件随机场模型对句子进行标注，最终根据标注序列上不同特征标签集的意义识别出人名。

2 条件随机场

在基于统计的标注方法中，条件随机场(Conditional Random Field, CRF)模型具有很好的效果，其模型思想主要来源于最大熵模型，但又不存在最大熵模型的数据稀疏问题，同时也无需对数据进行不必要的独立性假设，因而也优于隐马尔科夫模型（HMM）。条件随机场是一种用来标记和切分序列化数据的无向图模型，用于标记状态序列的CRF通常采用如图1的一阶链式结构。

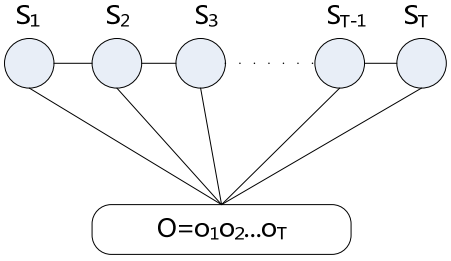

图 1 一阶链式结构

设 $O=\{o_1,o_2,...,o_T\}$ 为观察序列，例如有待标注词位的音节字序列 $S=\{s_1,s_2,...,s_T\}$ 表示被预测的状态序列，每一个状态均与一个词位标记(例如词首 B、黏写形式词尾 E’)相关联。这样，在一个观察序列给定的情况下，参数为 $\Lambda=\{\lambda_1,\lambda_2,...,\lambda_K\}$ 的一阶链式 CRF 模型的状态序列的条件概率为：

$$p(S|O)=\frac{1}{Z(o)}\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k f_k(s_{t-1},s_t,o,t)\right) \tag{1}$$

其中，$Z(o)$ 是归一化因子。$f_k(s_{t-1},s_t,o,t)$ 为特征函数，整合了观察序列和相应标记序列在 $t-1$ 和 t 时刻的转移特征以及 t 时刻整个观察序列和标记的状态特征，通常是一个二值表征函数。λ_k 是一个需要从训练数据中学习的参数，并为相应特征函数 $f_k(s_{t-1},s_t,o,t)$ 的权重，取值范围是 $(-\infty,+\infty)$。

3 特征模板的选择

3．1 特征标签的选择

在条件随机场模型中，最重要的问题是如何根据不同的任务选择合适的特征标签集。标签有基于字一级的粒度和词一级的粒度。考虑到在词一级上进行识别，需要以准确的分词结果为基础，如果分词有歧义反而会干扰识别的过程和结果，所以本文采用了基于字一级的特征标签[6, 7]。

本文中标签集的设定依据是某个字词在人名构成中所起的不同作用，如名字、前缀、后

缀、上文、下文等。与汉族人名类似，藏族人名的上下文用词都比较集中，有很强的规律性，同时人名的前缀与后缀的用词范围也很有限[8]。藏族人名与汉族人名不同的地方是，藏族人名没有严格意义上的姓氏，旧西藏贵族和宗教界人士名字中的家族名、寺庙名等被用作类似汉族的姓，大对数普通人的名字中没有“姓”的部分。藏族人名的长度也不固定，大多数为 2-4 个音节，但宗教界上层人士的名字往往很长，有 6 个音节、10 个音节、最长可达 26 个音节。同时，藏文中的格标记也可能以黏写形式附着在人名之后，使得藏族人名的自动识别更加困难。根据藏族人名的具体特点，本文设定的标签集如表 1 所示：

表 1 特征标签集

标签	意义	举例
B	名字首字	བཀྲ/B$^{ཤིས}$/E
M	名字中间字	ཆོས/B^{བཟང}/M$^{དོན}$/M$^{གྲུབ}$/E
E	名字末字	ཆོས/B^{བཟང}/E
E’	带黏着形式的名字末字	པད/B$^{མའི}$/E’
S	单字名字	ཡིའི/S^{ཡཁས}/U
S’	带黏着形式的单字名字	སྒོམ/P$^{དཔོན}$/P པའི/S' དང/A
P	名字前缀	ཀློའ/P$^{རིག}$/P$^{མིག}$/B^{དམར}/E^{ཡཁས}/U
U	名字后缀	རྒན/P$^{རོང}$/B^{ཐ}/E^{ཡཁས}/U
F	名字上文	ང/F^{ས}/F$^{རོ}$/B$^{རྗེ}$/E^{ལ}/Aསྒོམ/Nསྒོ ང/N$^{ཐེད}$/N$^{ཀྱི}$/N$^{ཡིན}$/N
A	名字下文	རྒན/P^{ཕང}/S^{ཡཁས}/z$^{ཀྱིས}$/A
C	名字间连接字	སྒོལ/B^{དཀར}/E^{ལས}/C$^{ཀྱ}$/B$^{མོ}$/E
N	无关字	རྣམ/B$^{རྒོལ}$/E$^{ཀྱ}$/A^{ས}/A^{ར}/A$^{བསྟད}$ /A

以下为根据标签集标注的例子：

བོད/N^{ལ}/N^{སངས}/N$^{རྒྱས}$/N$^{ཆོས}$/N$^{ལུགས}$/N^{ས}/N$^{བྱུང}$/N$^{གོང}$/F$^{རྒྱལ}$/F$^{པོ}$/P$^{གཤེན}$/B$^{ཁྲི}$/M^{བཙན}/M$^{པོ}$/E^{རས}/A$^{རྒྱལ}$/P$^{པོ}$/P ཡུ/B$^{ཙོ}$/M^{ག}/M$^{རེ}$/M$^{གཤེན}$/M^{བཙན}/E$^{གྱི}$/A^{བར}/N$^{རྒྱལ}$/N^{རབས}/N$^{ཉི}$/N$^{ཤུ}$/N^{ཙ}/N$^{བདུན}$/N$^{རིང}$/N$^{བོད}$/N$^{ཀྱི}$/N^{ཆབ}$^{སྲིད}$/N$^{དེ}$/N$^{སྐྱ}$

ང་/N།/Nསྡེཀྵ/N།/Nབོན་/Nཀསུས་/Nཐྱིས་/Nབསྒྲངས་/Nཞེས་/Nཁསལ་/Nད་/Nདེ/Nར་Nཁཱི་/Nབཟང་/Nཅ/N།/N

3.2 特征模板的设定

条件随机场模型对特征集的使用是通过固定格式的特征模板来定义的，通过特征模板在一个上下文环境（窗口）中使用特征。上下文特征窗口扩展得越大，越能较好地观察到当前字与上下文的关系，更好地发现文本中的长距离依赖，提高识别准确率。但是窗口越大，模型的训练时间就越长，影响模型的整体性能。考虑到藏文人名与其上下文的联系及特征标签集中各成分的相互关系，将窗口长度设为 5，可以在训练时间和识别效果上达到一个平衡。

条件随机场模型常用的特征模板如下表所示，表 1 中的 U01，U02 指的是特征的序号，%x[0，0]指的是当前字的一元特征(Unigram)，%x[-1，0]/%x[1，0]指的是前一个字和后一个字组成的二元特征组(Bigram)，依此类推。

表 2 常用特征模板

模板类型	特征模板
Unigram	U00:%x[-2,0] U01:%x[-1,0] U02:%x[0,0] U03:%x[1,0] U04:%x[2,0]
Bigram	U05:%x[-2,0]/%x[-1,0] U06:%x[-1,0]/%x[0,0] U07:%x[0,0]/%x[1,0] U08:%x[1,0]/%x[2,0] U09:%x[-1,0]/%x[1,0]

以具体例子来说明：

ང་/Nབོད་/Fརིགས་/Fདཀི་/Pཆོན་/Pདོན་/Bགྲུབ་/Eཡཐས་/Uཅཚལ་/Aདགོས་/Nཡོད་/N།/N

假设我们采用特征模板（U02:%x[0,0]，U07:%x[0,0]/%x[1,0]）来处理，则从第一个音节字“ང”观察到的特征为（U02:（“ང”，N），U07:（“ངབོད་”），N），从第二个音节字“བོད་”观察到的特征为（U02:（“བོད་”，F），U07:（“བོད་རིགས་”），F）。

4 实验结果与分析

本实验的训练集是从社科院民族所建立的人工分词平衡语料中，抽取出的大约4万余字的语料，共包含藏文人名1951个。为了确保语料的准确性，全部经过人工反复校对。实验本身分为封闭测试与开放测试两部分。封闭测试随机抽取了训练集中10%的语料做为测试语料，开放测试则选取了中学藏文教材及藏译本《水浒传》节选语料做为测试语料。所有语料的训练与测试均采用CRF++开源程序包，实验结果见表3。

表 3 实验结果

	准确率（%）	召回率（%）	F1 值（%）
封闭测试	91.75	97.02	94.31
开放测试1	90.38	92.16	91.26
开放测试2	82.60	86.96	84.7

从实验结果看，封闭测试的效果最好，F值达到了94.31%，召回率达到了97.02%，原因在于其测试语料的环境与训练语料相同，训练得到的模型可以较好地适用于相同语境的文

本。开放测试1的结果略低于封闭测试结果，但基本与之相当，是因为开放测试 1 的语料同训练语料一样，都属于现代书面藏语的语料，语境较为相似，与训练得到的模型能较好匹配。开放测试2的结果较前两个测试有明显降低，原因在于开放测试2使用的语料为藏译本《水浒传》节选，测试语境明显区别于训练语料，导致了测试结果与封闭测试有一定差距。

除此之外，藏文人名的复杂性也导致了开放测试结果不理想，主要体现在以下几个方面：

一、藏文人名不但包括藏族人名，还存在大量的藏译人名。其中藏译汉族人名最常见，也有少量的外国人名及其他民族的音译人名，如ཞེའོ་ཏིང་“小林”、ཡུའེ་ཏན་ཧྲུན་“约翰逊”、ཨ་ཕན་ཏི“阿凡提”等，在识别处理时，往往由于训练数据的稀疏导致一些人名模式难以识别。

二、藏族人名在长度上有较大的差别，有单个字的，也有多达数十个字的，部分人名前缀很长，也造成识别困难。

三、带有黏写形式的人名能否正确识别依赖于黏写形式的正确识别、切分。比如ཨྲག་པ“拉巴”可能带属格黏写形式构成ཨྲག་པའི་ཨ་ཅག“拉巴的姐姐”；同时，还有可能遇到黏写形式需要还原的情况，如文本形式ཆོ་དགས“次噶”，实际应是ཆོ་དགའ“次噶”+ས“施格”。如果对黏写形式的判断发生错误，那么人名识别也会遇到问题。

四、藏族人名与普通实词同形，在格标记缺失情况下，导致识别错误。如དཔལ་འབྱོར可以是人名“班觉”，也可能是词“经济”，在下面例句中是人名，但由于没有前后缀，识别结果错误。

ང་/rhཡོང་/voདུས་/ntདཔལ་འབྱོར་/nhསྒོག་ཁང་/ngརྡོ་/nlལ་/wlའདུག/ve

对于前两种错误，我们将考虑分类识别的方法，先通过规则对人名进行粗分类，进而对不同类别的人名进行识别。第三种错误涉及到黏写形式的识别问题，需要针对黏写形式的识别错误改进特征标签集。第四种错误很难通过上下文特征信息来判别，因而处理难度极大，暂时没有特别有效的处理方法，还需要进一步的研究。

5 结论

藏文人名的识别是藏文信息处理的重要工作之一。本文基于条件随机场模型，给出了藏文人名识别的一种方法。因为条件随机场模型能实现上下文信息的提取，发现文本中的长距离依赖关系，故在识别藏文人名的任务中能获得满意的效果。不过，在训练语料和测试语料环境相差较大的情况下，例如开放环境，训练语料规模和文本领域尚需大规模提高和扩展。

另外，特征标签集的设定也是一个关键的环节。在本文中使用了名字边界、前后缀、上下文等特征，但未能充分发掘名字与普通词语同形这一极易导致歧义的现象的特征，成为导致识别效果不十分理想的因素之一。在今后的工作中，可以考虑对特征标签集进行调整，同时优化特征模板集，期望进一步提高识别效果。

参考文献

[1] 项保，张国喜. 汉藏机器翻译中汉族人名翻译问题探讨[J]. 青海师范大学学报, 2011, 4.

[2] 吕雅娟，赵铁军，等. 基于分解与动态规划策略的汉语未登录词识别[J]. 中文信息学报, 2001, 15(1)：28-33.

[3] 毛婷婷，李丽双，黄德根. 基于混合模型的中国人名自动识别[J]. 中文信息学报，2007，21(2)：22-28.
[4] 丁伟伟，常宝宝. 基于语义组块分析的汉语语义角色标注[J]. 中文信息学报，2009，23(5)：53-61.
[5] 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，2010，11(2)：113-115.
[6] 邱莎，段玻，申浩如，等. 基于条件随机场的中文人名识别研究[J]. 昆明学院学报，2011，33(6)：64-66.
[7] 唐钊. 条件随机场模型在中文人名识别中的研究与实现[J]. 现代计算机，2012，7：3-7.
[8] 张华平，刘群. 基于角色标注的中国人名自动识别研究[J].计算机学报，2004，27(1)：86-91.

作者联系方式：康才畯 地址：北京市朝阳区朝阳路十里堡甲 3 号院 8 号楼 6A 邮编：100025
电话：13810104955 电子邮箱：kang.caijun@gmail.com