文章编号：1003-0077(2019)08-0053-07

基于置信度的藏文人名识别的主动学习模型研究

王志娟[1,2]，刘飞飞[3]，赵小兵[1,2]，宋 伟[1]

（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少数民族语言中心，北京 100081；
3. 好未来教育科技集团，北京 100080）

摘 要：训练语料的标注成本是资源稀缺语言处理研究面临的一个重要问题，通过主动学习(active learning)方法可以选择信息量大、无冗余的语料供人工标注，进而大大降低语料标注成本。该文基于 CRF 模型给出的标注置信度提出了四种主动学习方法，并通过实验确定了这四种主动学习方法的相关参数。实验显示：选择置信度低于 0.7 的语料进行人工标注，直到新旧模型标注结果的差异度小于 0.01%时，仅需 6 轮迭代；人工标注 3.2MB 的语料，藏文人名识别的 F 值可以达到 88%，若要达到该识别效果，基于 CRF 的监督式学习模型需要标注约 10MB 的语料，该主动学习方法降低了约 66%的语料标注规模。

关键词：藏文人名识别；主动学习；置信度

中图分类号：TP391　　**文献标识码**：A

Confidence Based Active Learning Model for Tibetan Person Name Recognition

WANG Zhijuan[1,2], LIU Feifei[3], ZHAO Xiaobing[1,2], SONG Wei[1]

(1. School of Electronics Engineering, Minzu University of China, Beijing 100081, China;
2. National Language Resource Monitoring & Research Center of Minority Languages, Beijing 100081, China;
3. Tomorrow Advancing Life Education Group, Beijing 100080, China)

Abstract: To alleviate the issue of labeling cost of training data for low resource languages, the active learning is a promising method by selecting the informative data without redundancy. Four active learning methods based on the confidence are proposed, with the parameters decided empirically. The experimental results: selecting the data with confidence below 0.7 and 6 iteration of labeling with up to 3.2MB training data, we can achieve 0.88 F-measure for Tibetan name recognition. Compare with the 10MB training data for CRF model to achieve the same performance (with no more than 0.01% difference), the active learning approach reduces the annotation scale by 66%.

Keywords: Tibetan person name recognition; active learning; confidence

0 引言

命名实体识别(named entity recognition, NER)作为信息抽取的子任务，是自然语言处理任务的基础环节，是信息检索、知识图谱等研究的基础。经过多年发展，命名实体识别研究覆盖了英语、汉语、印地语、阿拉伯语、日语、西班牙语等多种语言。

命名实体识别的主要方法有规则、机器学习和深度学习三类[1-2]。根据标注语料的规模，机器学习又可以分为监督式机器学习(训练语料全部标注)、半监督式机器学习(训练语料部分标注)和无监督式机器学习(无标注语料)三种，其中，基于大规模标注语料的监督式学习方法的命名实体识别性能优于半监督和无监督方法，是常用的命名实体识别方法。

对于资源稀缺语言而言，大规模、高质量标注语

收稿日期：2018-07-20　**定稿日期**：2018-08-10

基金项目：国家自然科学基金(61331013, 61501529)

料意味着更高的时间、人力和资金成本，因此如何以较低成本获取大规模、高质量标注语料是资源稀缺语言监督式学习方法所要解决的关键问题之一。另外，如何最大限度地避免重复标注工作、提高标注效率也是语料标注工作要解决的问题。例如，在进行藏文新闻语料的命名实体人工标注时发现，1 000个人民网藏文网页中有3 268个人名，其中，“习近平(ཞི་བཞིན་ཕིང་)”的出现次数高达502次，占所有标注人名总数的15.4%。因此，为了降低语料的标注成本，应该选择那些不确定性高、信息量大、没有冗余的语料进行人工标注。

主动学习是机器学习的一个子领域，其主要工作是有针对性地选择一些信息量大的语料进行人工标注，进而通过较少的标注语料实现较好的模型学习效果，从而最大限度地降低语料标注成本[3-4]。目前主动学习方法已经成功应用于许多自然语言处理任务，例如，文本分类[5]、词性标记[6]、词义消歧[7]、自动翻译[8]、命名实体识别[9-12]等。

本文提出了一种基于置信度的藏文人名识别的主动学习模型，该模型用约33%的人工标注语料就可达到监督式学习模型的藏文人名识别效果。

本文的主要内容安排如下：首先介绍了藏文人名识别的研究现状、面临的困难以及主动学习的原理，其次介绍了基于置信度的藏文人名识别的主动学习模型，然后是实验部分，最后是结论和展望。

1 相关工作

首先介绍藏文人名的识别研究现状，然后介绍主动学习的原理。

1.1 藏文人名识别现状

早期的藏文命名实体识别的研究主要采用基于规则的方法，Yu等[13]提出利用格助词、边界特征、词典等识别藏文命名实体的方法，Sun等[14]提出基于多特征的藏族人名识别方法，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实现藏文人名的识别。

2014年之后，藏文命名实体的识别方法开始以基于监督式机器学习方法为主。加羊吉等[15]提出了最大熵和条件随机场相融合的藏文人名识别方法；华却才让等[16]提出基于感知机的藏文命名实体识别；康才峻等[17]提出了基于条件随机场的藏文人名识别方法；2017年，珠杰等[18]基于条件随机场以及触发词、虚词、人名词典、人名后缀等特征的不同优化组合实现了藏文人名识别。

目前藏文人名识别研究已经取得了较好的识别效果，不过还存在音译人名及与普通名词同形的藏文人名识别效果不理想的问题[15]。这些问题往往是由于训练语料覆盖面不够所致，而藏文是一种资源稀缺语言，大量语料的标注将需要更高的人力、物力和财力成本，对此本文提出了一种基于置信度的主动学习方法，该方法将选择那些信息量大、无冗余的语料进行人工标注，进而达到降低语料标注成本的目的。

1.2 主动学习原理

主动学习是半监督机器学习的特例，该方法主要用于构造有效训练集，由于训练集中通常包含大量的冗余样本，主动学习方法从大量未标注语料中通过一定的选择策略选择一定数量的语料进行人工标注，从而降低语料标注成本[3]。

主动学习方法可以由以式(1)所示的五个组件进行建模[19]。

$$A=(C,L,S,Q,U) \tag{1}$$

其中，C为分类器，L为已标注的训练语料；S为语料标注人员；Q为选择策略，用于从未标注的语料中选择信息量大的语料供人工标注；U为整个未标注语料。

主动学习方法主要分为两个阶段：第一阶段为初始化阶段，利用已标注的语料建立一个初始分类器模型；第二阶段为迭代选择阶段，利用第一阶段构建的分离器标注未标注语料U，并按照某种选择策略Q从U中选取一定数量的语料交给标注者S进行标注，然后，将人工标注结果添加到已标注语料L中，重新训练分类器，直至满足停止标准为止[20]。

1.3 主动学习在命名实体识别方面的应用

目前，主动学习方法已被应用于命名实体识别任务中，Shen等[9]提出了一种基于多特征的主动学习方法，该方法将信息性、代表性、多样性三种特征

进行表示量化，通过融合这三种特征的选择策略减少了人工标注成本。实验显示：在保证识别效果的前提下，该方法可以减少约 80% 的语料标注量。Yao 等[11]提出了基于信息密度的选择策略，该方法仅利用约 1 万个标注句子就实现了人工标注约 13 万句子的效果。

针对藏文人名识别中由于训练语料稀疏导致的识别效果不理想的问题，理论上可以通过增加训练语料规模解决。本文基于不确定主动学习算法，利用条件随机场作为藏文人名识别模型，选择模型标注结果中置信度较低的语料进行人工标注，进而可以在保证识别效果的前提下，大大减少语料的人工标注成本。

2　基于主动学习的藏文人名识别模型

藏文人名识别的主动学习过程，如图 1 所示。

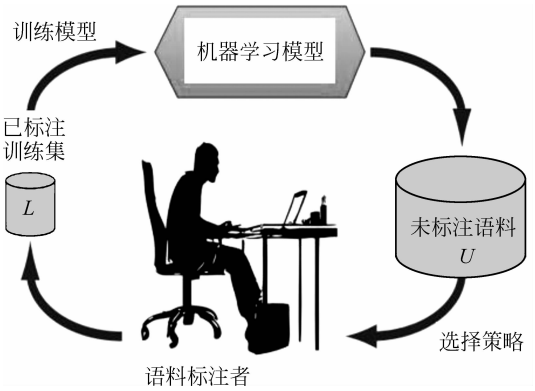

图 1　主动学习原理

首先给定少量人工标注语料 L 和大量未标注语料 U。然后按以下步骤训练基于主动学习的藏文人名识别模型。

第一步：利用人工标注语料 L 训练一个基于 CRF 的藏文人名识别模型 M_L。

第二步：用 M_L 去标注大量未标注语料。

第三步：在标注结果中按一定的选择策略选择若干不确定性高、信息量大的语料，交给人工标注。

第四步：将人工新标注的语料添加到已标注语料 L 中，同时将其从未标注语料 U 中删除。

第五步：判读是否满足主动学习结束条件，若满足，则结束；若不满足，则重复步骤一到五，直到满足主动学习结束条件。

因此，对于基于主动学习的藏文人名识别模型而言，选择策略和停止策略的设计至关重要，下面基于置信度和新旧模型标注结果的差异度分别介绍两种选择策略和两种停止策略。

2.1　基于置信度的选择策略

本文基于 CRF 模型识别藏文人名，对于给定的输入序列 X，其标注结果为 Y 的条件概率为 $P(Y|X)$，该结果的范围为[0,1]，0 表示对标注结果没有信心，1 表示完全确认标注结果[20]，如式(2)、式(3)所示。

$$P(Y \mid X) = \frac{1}{Z(x)} \prod_{t=1}^{T} \varphi_t(Y_t, Y_{t-1}, X_t) \tag{2}$$

$$\varphi_t(Y_t, Y_{t-1}, X_t) = \exp\Big\{ \sum_{k=1}^{K} \theta_k f_k(Y_t, Y_{t-1}, X_t) \Big\}) \tag{3}$$

标注结果的置信度计算方法如式(4)所示。

$$\text{Confidence}(X) = P(Y \mid X) \tag{4}$$

本文基于句子的置信度选择需要人工标注的语料，具体选择策略有两种。

(1) 选择策略 1

该选择策略的基本思想是每次迭代选择置信度最低的前 m 个句子进行人工标注，因此每次选择的句子数 m 是固定的。

(2) 选择策略 2

该选择策略的基本思想是每次迭代选择置信度低于某个阈值的 n 个句子进行人工标注。该方法每次迭代选择的句子数可能不一样，随着迭代次数的增加，每次选择的句子数 n 会越来越少。

2.2　停止策略

本文提出了两种停止策略。

(1) 停止策略 1：基于置信度的停止策略。

该停止策略的基本思想是当所有待选语料的置信度均高于设定的阈值 α 时，主动学习停止。

(2) 停止策略 2：基于差异度的停止策略。

该停止策略的基本思想是将新、旧模型标注结果的差异度 β 作为停止依据，新、旧模型标注结果的差异度越小，说明新、旧模型性能的差异越小，当二者的差异小于一个足够小的数时，主动学习过程结束。

新、旧模型的差异度计算方法如下：

假定对音节 x_i，新模型的标注结果为 L_{x_i}，旧模型的标注结果为 L'_{x_i}，此音节的标注差异如式(5)所示。

$$\mathrm{diff}(x_i)=\begin{cases}0, & \text{如果 } L_{x_i}=L'_{x_i}\\ 1, & \text{其他}\end{cases} \tag{5}$$

新旧模型的标注差异度计算如式(6)所示。

$$\mathrm{Diff}=\frac{\sum_{i=1}^{n}\mathrm{diff}(x_i)}{n}*100\% \tag{6}$$

其中，$\mathrm{diff}(x_i)$表示第 i 个音节的标注差异情况，n 表示模型标注的音节总数。理论上而言，当新、旧模型的差异度 β 为 0 或者小于一个非常小的数时，表示新、旧模型标注结果基本一致，主动学习可以停止。

2.3 基于置信度的主动学习方法

基于以上提出的选择和停止策略，有以下 4 种主动学习方法。

方法 1 选择策略 1＋停止策略 1

该主动学习方法每次迭代选择固定数量(m个)的句子供人工标注，直到待选语料的句子置信度均高于设定置信度阈值 α_1 为止。

方法 2 选择策略 1＋停止策略 2

该主动学习方法每次迭代选择固定数量(m)的句子供人工标注，直到新、旧模型的标注结果的差异度小于设定阈值(β_1)为止。

方法 3 选择策略 2＋停止策略 1

该主动学习方法每次迭代选择置信度低于给定阈值(n)的若干句子供人工标注，直到待选语料的置信度均高于设定阈值 α_2 为止。

方法 4 选择策略 2＋停止策略 2

该主动学习方法每次迭代选择置信度低于给定阈值(n)的若干个句子供人工标注，直到新、旧模型的标注结果的差异度小于设定阈值(β_2)为止。

以上参数均由实验确定。

3 实验

首先介绍实验方案，然后根据实验确定主动学习方法 1～4 中的各个参数，从标注效果、标注语料量和迭代次数三方面分析这四种主动学习方法的性能，最后比较主动学习方法和监督式学习的效果。

3.1 实验设计

本实验语料来自人民网、藏语广播网、阿坝新闻网的藏语版，语料覆盖新闻、政治、宗教、文化等多个领域，不仅包含大量藏族人名，还包含大量译名。实验语料一共 1 500 个文本，其中训练语料 1 360 个文本(人工标注语料 100 个文本、未标注语料 1 260 个文本)、测试语料 140 个文本，语料基本情况如表 1 所示。

表 1 实验语料基本情况

语料		文本数	规模(MB)
训练语料	已标注语料	100	0.2
	待标注语料	1 260	9.8
测试语料		140	1
共计		1 500	11

3.2 主动学习方法的参数确定

3.2.1 方法 1 的参数确定

表 2 所示为当选择策略为每次迭代选择置信度最低的 50、100、150、200、250 句，停止策略为标注结果的置信度为 0.5～0.9 时藏文人名识别效果、主动学习迭代次数及语料标注规模。

由表 2 可见，选择 $m=50$、$\alpha1=0.8$ 时，藏文人名识别的 F 值可达到 88.3%，主动学习迭代次数为 63 次，语料标注规模为 2.57 MB。

表 2 方法 1 不同参数的藏文人名识别效果

m	α_1/%	F_1/%	迭代次数	标注语料规模/MB
50	0.5	87.0	31	1.47
	0.6	87.9	42	1.87
	0.7	87.5	50	2.15
	0.8	88.3	63	2.57
	0.9	88.3	85	3.39
100	0.5	87.6	15	1.49
	0.6	88.1	19	1.77
	0.7	88.2	25	2.16
	0.8	88.1	31	2.57
	0.9	87.9	41	3.33
150	0.5	86.4	9	1.40
	0.6	87.7	13	1.83
	0.7	87.9	17	2.20
	0.8	87.9	21	2.61
	0.9	88.2	28	3.45

续表

m	α_1/%	F_1/%	迭代次数	标注语料规模/MB
200	0.5	87.5	7	1.65
	0.6	87.6	9	2.05
	0.7	88.0	13	2.85
	0.8	88.0	15	3.25
	0.9	87.9	21	4.45
250	0.5	85.9	5	1.35
	0.6	87.2	7	1.71
	0.7	87.6	10	2.26
	0.8	88.1	13	2.71
	0.9	88.2	17	3.43

3.2.2　方法 2 的参数确定

表 3 所示为当选择策略为每次迭代选择置信度最低的 50、100、150、200、250 句，停止策略为标注结果的差异度为 0.02%、0.01%、0.005%时藏文人名识别效果、主动学习迭代次数及语料标注规模。

由表 3 可见：选择 $m=250$、$\beta_1=0.01\%$时，藏文人名识别的 F_1 值可达到 88.1%，主动学习迭代次数为 13 次，语料标注规模为 2.71 MB。

表 3　方法 2 不同参数的藏文人名识别效果

m	β_1/%	F_1/%	迭代次数	标注语料规模/MB
50	0.02	84.5	15	0.88
	0.01	86.4	27	1.32
	0.005	87.9	47	2.03
100	0.02	86.0	13	1.35
	0.01	87.4	21	1.90
	0.005	87.9	37	3.00
150	0.02	87.4	11	1.62
	0.01	87.9	21	2.61
	0.005	—	—	—
200	0.02	87.7	11	2.00
	0.01	87.9	17	2.81
	0.005	—	—	—
250	0.02	87.9	11	2.34
	0.01	88.1	13	2.71
	0.005	—	—	—

3.2.3　方法 3 参数的确定

由于方法 3 的选择策略 n 和停止策略 α_2 均基于置信度，因此二者的取值只能相等。假定选择策略和停止策略同等重要，令 $n=\alpha_2=0.5$，此时的标注效果、标注规模及迭代次数如表 4 所示，可见，基于该主动学习方法，藏文人名识别的 F_1 值为 86.9%，主动学习迭代次数为 18 次，语料标注规模为 2.05MB。

表 4　方法 3 的藏文人名识别效果

n	α_2	F_1/%	迭代次数	标注语料规模/MB
0.5		86.9	18	2.05

3.2.4　方法 4 参数确定

表 4 所示为当选择策略的置信度阈值为 0.4～0.7，停止策略的差异度为 0.02%、0.01%、0.005%时对应的藏文人名识别效果、主动学习迭代次数及语料标注规模。

表 5　方法 4 不同参数的藏文人名识别效果

n	β_2/%	F_1/%	迭代次数	标注语料规模/MB
0.4	0.02	85.5	5	1.34
	0.01	85.6	7	1.37
	0.005	86.3	13	1.41
0.5	0.02	86.6	5	1.97
	0.01	87.1	7	2.02
	0.005	87.1	7	2.02
0.6	0.02	87.6	5	2.25
	0.01	87.6	5	2.56
	0.005	87.3	7	2.60
0.7	0.02	79.8	5	3.22
	0.01	88.0	6	3.23
	0.005	87.8	8	3.24

由表 5 可见：综合考虑识别效果、语料标注规模及迭代次数，选择 $n=0.7$、$\beta_2=0.01\%$时，藏文人名识别的 F_1 值可达到 88.0%，此时，主动学习迭代次数为 6 次，语料标注规模为 3.23 MB。

3.2.5　监督式学习方法与主动学习方法对比

表 6 是基于不同标注语料规模的监督式学习模型的藏文人名识别效果[21]。可见，当所有训练语料(10.26 MB)均已人工标注的条件下，藏文人名识别

的 F_1 值最高可达 88.3%。

表 6 语料规模对藏文人名识别效果的影响(基于 CRF)

标注语料规模		F_1/%	标注语料规模		F_1/%
文本数	规模/MB		文本数	规模/MB	
100	0.20	38.1	800	5.60	78.7
200	0.89	63.4	900	6.29	81.7
300	1.84	69.2	1 000	6.82	82.0
400	2.83	72.4	1 100	7.66	83.3
500	3.54	76.0	1 200	8.63	85.1
600	4.18	76.1	1 300	9.81	86.7
700	4.97	77.3	1 360	10.26	88.3

表 7 所示为藏文人名识别的监督式学习方法和主动学习方法的对比情况。

表 7 监督式学习方法与主动学习方法对比

方法	选择策略	停止策略	F_1/%	迭代次数	语料标注规模/MB
方法 1	m=50	α_1=0.8	88.3	63	2.57
方法 2	m=250	β_1=0.01%	88.1	13	2.71
方法 3	n=0.5	α_2=0.5	86.9	18	2.05
方法 4	n=0.7	β_2=0.01%	88.0	6	3.23
监督式学习			88.3	1	10.26

由表 7 可见:

(1) 主动学习方法可以基于较少的标注语料达到基于较多标注语料的监督式学习方法的识别效果。本文提出的主动学习方法 1、2、4 仅用约 30% 的人工标注语料就达到了基于 10 MB 标注语料的监督式学习方法的藏文人名识别效果。

(2) 主动学习方法的效果取决于选择策略和停止策略的设计,主动学习方法的评价指标除了 F_1 值,还有循环迭代次数以及语料标注量。

主动学习方法 1 具有最好的识别效果(88.3%)以及最少的语料标注量(2.57 MB),但是方法 1 的循环迭代次数高达 63 次,语料标注周期过长;

主动学习方法 2 具有较好的识别效果(88.1%)以及较少的语料标注量(2.71 MB),但方法 2 的循环迭代次数为 13 次,语料标注周期相对也过长;

主动学习方法 4 所需的时间迭代次数最少,藏文人名的识别效果略低于方法 1(方法 4 的 F_1 值约为 88.0%),但方法 4 的语料标注量最大(约 3.23 MB)。

综合识别效果、迭代次数以及语料标注规模三个因素,我们选择方法 4 作为藏文人名的主动学习模型。

4 总结与展望

语料标注成本是资源稀缺语言自然处理研究面临的问题之一,主动学习方法通过选择一些信息大、不确定性高、无冗余的语料进行人工标注,进而在保证效果的前提下,大大降低语料标注成本。本文基于置信度提出了四种主动学习方法,实验证明:主动学习方法 4(每次迭代选择置信度低于 0.7 的句子进行人工标注,直到新、旧模型标注结果的差异度小于 0.01%)可用 3.23 MB 的标注语料、在最少的迭代次数近似达到监督式学习方法 10 MB 标注语料的效果,人工语料标注量降低了约 66%。

基于主动学习的藏文人名识别模型中,识别效果、迭代次数以及语料标注规模三个因素有的互为促进关系、有的互为制约关系,今后可以从这三因素的关系出发对选择策略和停止策略进行进一步优化设计,进而达到以最低的人力、时间成本获取大规模、高质量标注语料的目的。

参考文献

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

[2] 赵军. 命名实体识别、排歧和跨语言关联[J]. 中文信息学报, 2009, 23(2): 3-17.

[3] Settles B. Active learning literature survey [D]. University of Wisconsinmadison, 2009, 39(2): 127-131.

[4] Culotta A, Kristjansson T, Mccallum A, et al. Corrective feedback and persistent learning for information extraction[J]. Artificial Intelligence, 2006, 170 (14-15): 1101-1122.

[5] Hoi S C H, Jin R, Lyu M R. Large-scale text categorization by batch mode active learning[C]//Proceedings of the 15th International Conference on World Wide Web, ACM, 2006: 633-642.

[6] Ringger E, Mcclanahan P, Haertel R, et al. Active learning for part-of-speech tagging: Accelerating corpus annotation[C]//Proceedings of Linguistic Annotation Workshop. Association for Computational Linguistics, 2007: 101-108.

[7] Reichart R, Rappoport A. An ensemble method for selection of high quality parses[C]//Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, 2007: 408-415.

[8] Kuo J S, Li H, Yang Y K. Learning transliteration lexicons from the web[C]//Proceedings of International Conference on Computational Linguistics and the Meeting of the Association for Computational Linguistics, 2006: 1129-1136.

[9] Shen D, Zhang J, Su J, et al. Multi-criteria-based active learning for named entity recognition[C]//Proceedings of Meeting on Association for Computational Linguistics, 2004: 589-596.

[10] Chen Y, Lasko T A, Mei Q, et al. A study of active learning methods for named entity recognition in clinical text[J]. Journal of Biomedical Informatics, 2015, 58(C): 11-18.

[11] Yao L, Sun C, Li S, et al. CRF-based active learning for chinese named entity recognition[C]//Proceedings of 2009 IEEE International Conference on Systems, Man and Cybernetics, 2009: 1557-1561.

[12] Tran V C, Nguyen N T, Fujita H, et al. A combination of active learning and self-learning for named entity recognition on Twitter using conditional random fields [J]. Knowledge-Based Systems, 2017, 132: 179-187.

[13] Yu H, Jiang T, Ma N. Named entity recognition for Tibetan texts using case-auxiliary grammars [J]// Proceedings of International Multi Conference of Engineers and Computer Scientists. 2010, 2180(1).

[14] Sun Y, Yan X, Zhao X, et al. Research on automatic recognition of Tibetan personal names based on multi-features[C]//Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering. IEEE, 2010: 1-5.

[15] 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.

[16] 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.

[17] 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.

[18] 珠杰，李天瑞，刘胜久. 基于条件随机场的藏文人名识别技术研究[J]. 南京大学学报(自然科学)，2016，52(2)：289-299.

[19] 吴伟宁，刘扬，郭茂祖，等. 基于采样策略的主动学习算法研究进展[J]. 计算机研究与发展，2012，49(6)：1162-1173.

[20] Lafferty John D, McCallum, 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.

[21] 刘飞飞，王志娟. 基于层次特征的藏文人名识别研究[J/OL]. 计算机应用研究，2018(09)：1-7 [2018-05-14]. http://kns.cnki.net/kcms/detail/51.1196.TP.20170828.1023.066.html.

王志娟(1977—)，通信作者，博士，副教授，主要研究领域为自然语言处理。
E-mail: wangzj_muc@126.com

赵小兵(1967—)，博士，教授，主要研究领域为自然语言处理。
E-mail: nmzxb_cn@163.com

刘飞飞(1993—)，硕士，主要研究领域为自然语言处理。
E-mail: liufeifei_muc@163.com