文章编号：1003-0077（2011）00-0000-00

基于知识融合的 CRFs 藏文分词系统

洛桑嘎登[1]，杨媛媛[2]，赵小兵[3]

（1 中央民族大学信息工程学院，北京 100081

2 中央民族大学少数民族语言文学系，北京 100081

3 中央民族大学国家语言资源监测中心少数民族语言分中心，北京 100081）

摘要：藏文分词问题是藏文自然语言处理的基本问题之一，本文首先通过对 35.1M 的藏文语料进行标注之后，通过条件随机场模型对其进行训练，生成模型参数，再用模版对未分词的语料进行分词，针对基于条件随机场分词结果中存在的非藏文字符切分错误，藏文黏着词识别错误，停用词切分错误，未登录词切分错误等问题分别总结了规则，并对分词的结果利用规则进行再加工，得到最终的分词结果，开放实验表明该系统的正确率 96.11%，召回率 96.03%，F 值 96.06%。

关键词：藏文；分词；条件随机场；知识融合

中图分类号：TP391　　**文献标识码：**A

Research of Tibetan Automatic Word Segmentation Based on Conditional Random Field And Knowledge Fusion

Luobsang Karten[1],Yang Yuanyuan [2], Zhao Xiaobing[3]

(1 Shool of Information Engineering ,Minzu University of China, Beijing,China;

2 School of Chinese minority language and literature, MinZu university of China, BeiJing,China;

3Minzu University of China,National Language Resource Monitoring & Research Center of Minority Languages, Beijing,China)

Abstract: Tibetan word segmentation is one essential task in Tibetan language processing. In this paper, first tagging Tibetan corpus of 35.1M, training the tagged corpus with CRFs, generating model parameter, segmenting raw corpus with model, then summarizing rules for the errors produced in CRF segmentation results, such as segmentation errors of non-Tibetan characters, recognition error of Tibetan adhesion words, segmentation errors of stop words and unregistered words, and reprocessing the word segmentation results to get the final result, achieving an accuracy of 96.11%, recall rate of 96.03%, F score of 96.06%.

Key words: Tibetan;Word Segmentation;CRFs;Knowledge Fusion

1 引言

藏文自动分词可以看作是计算机自动辩识藏文文本字符流中的词，并在词与词之间加入明显的词切分标记符的过程[1]。藏文自动分词的主要目的是确定藏文信息处理处理的基本语言

* **收稿日期：** 2015 年 6 月 15　**定稿日期：**2015 年 8 月 10

基金项目：国家自然科学基金重点项目“跨语言社会舆情分析基础理论与关键技术研究”(61331013)

单位，为进一步开展藏文智能分析和处理做好前期准备工作。目前藏文分词技术的研究方法大体可以分成两类，一种是基于藏文自身的语法特点，首先将文本通过标点分成句子，其次通过格助词将句子分成组块，最后再对组块内部通过匹配等方法将词与词分开。另一种是基于统计的方法，将在中文分词中取得不错效果的统计自然语言的方法，例如隐马尔科夫，最大熵，条件随机场等移植到藏文自然言语处理过程中。

2 相关研究

藏语分词作为藏文信息处理中重要的基础工作，迄今为止已经有不少学者进行了研究。最早的关于藏文分词系统的研究可以追溯到 1997 年江荻进行了规则分词技术研究，提出藏语最大匹配算法、任意词和句尾词分词匹配校验等设计方案。1999 年，罗秉芬、江荻等从 12 万词条和 500 万字藏语真实文本语料分词的实践中归纳出了藏文计算机自动分词的 36 条基本规则，并提出了藏文分词的基本框架 [2]。同年扎西次仁基于 5000 多个常用词词表，利用最大匹配法和人工校对的方式实现了分词功能，但是由于词库和方法上的局限性，该系统仅仅具备演示效果，不具备实用性 [3]。2003 年，陈玉忠从藏文的语法接续规则出发，提出了基于格助词和接续特征的书面藏语自动分词方案 [4]，并依据该分词方案的总体设计思路，陈玉忠等阐述了书面藏语自动分词系统的具体实现过程 [5]。该方案在将藏文句子分块的过程中增加了藏文语法中接续规则，一定程度上提高了分词的准确性，但是无法切分的块，采取加标记但不切分的“谨慎”策略，并默认其属于未登录词。这样的做法显然对未登录词的识别不够精确。2009 年，才智杰设计了“班智达藏文分词系统”[6]，此系统分三步实现分词功能，首先将文本分成句字，再通过格助词将句子分成组块，块内再通过词典匹配切分成词，并对词典搜索算法进行了改进，既对词典进行按照词长排序，以提高搜索速度。但是此方法针对在分词中存在的歧义问题，没有给出合理的处理方法。

以上系统实现的技术思路主要是根据藏语中的接续特征[7]，字、词、句各级语言单位之间的自然切分标记，先利用字切分特征、字性库“认字”，再用标点符号、关联词“断句”，用格助词“分块”，最后通过词典匹配“认词”。该技术方案进一步发展为组块分词策略，即充分利用藏语丰富的句法形式标记，通过各类名物化标记、格标记、指代词、连词、动词语尾、构词词缀等形式标记构建不同的藏语句法组块类型，并建立相应的组块规则，分词时先根据形式标记和规则分块，然后在块内进行分词。在具体操作时采取最大正向匹配法、最大逆向匹配法或者是最大双向匹配法等不同的策略。

随着汉语分词开始使用各种统计机器学习模型，如隐马尔科夫模型、最大熵马尔科夫模型、条件随机场模型等，基于统计的藏语分词研究成果也逐渐多起来。2011 年，史晓东、卢亚军率先把统计方法引入藏语分词研究，他们开发的央金藏文分词系统把汉语分词系统 Segtag 的技术移植到藏语分词中，实现了藏语的分词标注一体化 [8]。该系统主要采用的隐马尔科夫模型，使用了约 2.7M 文本作为训练语料，其分词结果 F 值为 91.115%。 2012 年，刘汇丹等在研究分析了藏文分词中的格助词分块、临界词识别、词频统计、交集型歧义检测和消歧等问题之后，设计实现了一个藏文分词系统 SegT[9]，该系统采用双向切分检测交集型歧义字段进行消歧处理，在系统分词的正确率上得到了很大的提升。此外，江荻 [10]、羊毛卓么 [11]、扎西加 [12] 等学者还对藏文词语词形变体识别规则、词组结构以及词性标注等方面进行了研究，总体上推进了藏文分词以及文本分析研究的进展。

3 基于条件随机场的藏文自动分词

3.1 条件随机场相关介绍

条件随机场(Conditional Random Fields，CRFs)是一种基于统计的序列标记识别模型，它由 John Lafferty ， Andrew McCallum 和 Femando Peerira 在 2001 年首次提出 [10][11]。它是一种无向图模型，对于指定的节点输入值，它能够计算指定的节点输出值上的条件概率，其训练目标是使得条件概率最大化 [11]。线性链是 CRFs 中常见的特定图结构之一，它由指定

的输出节点顺序链接而成。一个线性链与一个有限状态机相对应，可用于解决序列数据的标注问题。下面，如果不加说明，CRFs 均指线性的 CRFs。用 x=（ x1，x2，…，xn）表示要进行标注的数据序列，y =（ y1，y2，…，yn）表示对应的结果序列。例如对于藏文分词任务，x 可以表示一个藏文句子 x =（“ཕོ་བྲང་པོ་ཏ་ལའི་རྒྱབ་ཀྱི་རྫོང་རྒྱབ་ཀླུ་ཁང་ནང་མེ་ཏོག་ཁྲ་ཆིལ་ཆིལ་དུ་བཞད་ཅིང་ཡུལ་ལྗོངས་ལྷ་ན་སྡུག་པ་ཞིག་ཡོད། ”），y 则表示该句子中每个音节所在位置的序列 y =（ B,E,B,M,Eg,S,S,B,M,M,E,S,B,E,B,M,E,S,S,S,B,E,B,M,E,S,S,S)。

对于(X，Y），C 由局部特征向量 f 和对应的权重向量 λ 确定。对于输入数据序列 x 和标注结果序列 y，条件随机场 C 的全局特征表示为：

$$\mathrm{F}(\mathrm{y},\mathrm{x}) = \sum_{i} \mathrm{f}(y,x,i) \quad (1)$$

其中 i 遍历输入数据序列的所有位置，f(y，x，i) 表示在 i 位置时各个特征组成的特征向量。于是，CRFs 定义的条件概率分布为：

$$P_{\lambda}(Y,X) = \frac{\exp[\lambda \bullet F(Y,X)]}{Z_{\lambda}(X)} \quad (2)$$

其中：

$$Z_{\lambda}(X) = \sum_{y} \exp[\lambda \bullet F(y,x)] \quad (3)$$

给定一个输入数据序列 X，标注的目标就是找出其对应的最可能的标注结果序列，即：

$$\overline{\mathrm{y}} = \arg_{y} \max p_{\lambda}(y \mid x) \quad (4)$$

由于 $Z_{\lambda}(X)$ 不依赖于 y，因此有:

$$\overline{\mathrm{y}} = \arg y \max p_{\lambda}(y \mid x) = \arg_{y} \max{}_{\lambda} \bullet F(y,x) \quad (5)$$

CRFs 模型的参数估计通常采用 L-BFGS 算法实现，CRFs 解码过程，也就是求解未知串标注的过程，需要搜索计算该串上的一个最大联合概率，解码过程采用 Viterbi 算法来完成。

CRFs 具有很强的推理能力，能够充分地利用上下文信息作为特征，还可以任意地添加其他外部特征，使得模型能够获取的信息非常丰富。CRF 模型没有隐马尔可夫模型(Hidden Markov Model，HMM)的强独立性假设条件，因此可以加入更多的文本信息特征；而且 CRFs 模型计算的是全局而非局部最优输出结点的条件概率，正因如此它解决了最大熵模型(Maximum EntroPy Model，MEM)的标记偏置问题。CRFs 模型能更容易的融合客观世界数据的真实特征，因此，此模型被广泛用于自然语言处理的很多领域。

3.2 基于 CRFs 的藏文分词

3.2.1 总体流程

如图 1 所示，我们对整个实验的流程做简单的陈述:

第一步 首先将从西藏新闻网、人民网藏语频道和青海藏语广播网爬取的语料经过预处理之后，通过词典匹配分词，再经过先后三次的人工校正，形成训练语料。将训练语料进行标注转换后，利用 CRFs 模型对转换后的语料进行训练，最终生成模型参数。

第二步 对来自新华网的语料，经过预处理之后，进行词典匹配分词，再经过先后三次的人工校正，形成测试语料。

第三步 通过测试语料反复测试结果，确定特征模板。

第四步 通过分析 CRFs 分词结果中的典型错误设计规则，在上一步识别的基础上，进行二

次识别，最终得到分词结果。

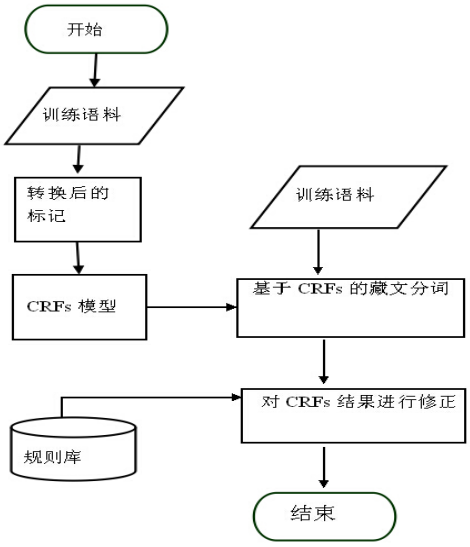


图 1 基于 CRF 和规则的藏文分词的流程图

3.2.2 藏文自动分词标注集的选择

我们首先定义条件随机场模型的训练所需要的标准集，标注集的目的是确定某个音节在藏文词的位置，以此确定某个藏文词的边界。而藏文文本存在其特有的黏着词，所以在对音节标注时，对于非单个音节构成的词的右边界和单音节构成的词需要区分是黏着形式还是非黏着形式，因此目前研究者在基于条件随机场的藏文分词的标注集的选择上分为两种方法，第一种是先标注黏着词，即在分词之前先通过二元标注集（Y/N）标注当前词是否为黏着形式，再进行分词 [12]。第二种方法是直接在标注音节位置的同时增加两个新的标签，即五元标注集（B，I，E，S，SS，ES）[13]。本文使用第二种方法。以藏文的每个音节为对象，标注集中主要定义了藏文词汇的开始音节、内部音节、结尾音节、黏着形式的单字和黏着形式的结尾音节，共 5 种类型，如表 1 所示。

表 1 藏文分词标注集

标注	含义
B	藏文词汇的首个音节
I	藏文词汇的中间音节
E	藏文词汇的结尾音节
Eg	带黏着形式的结尾音节
S	单个音节
Sg	带黏着形式的结尾音节

3.2.3 藏文自动分词特征集的选择

使用 CRFs 进行藏文分词的过程就是给定一个藏文句子 x = (x1，x2，…，xn) ，通过 Viterbi 解码算法找出其对应的每个音节的位置信息的结果序列 y = (y1，y2，…，yn) ，使得条件概率 P λ (y | x) 最大。而在基于 CRFs 的标注分类问题中，特征函数的选择通常起着关键性作用，特征选择的好坏直接决定着 CRFs 标注结果的优劣。CRFs 最大的优点之一就是特征的选择很灵活，根据要解决的问题，能够融入任意的特征。选择不同的特征，所得到的实验结果是不相同的。在本实验中，对于特征的选择，利用了词的上下文信息，这里所谓

的“上下文”可以看作是以当前词为基线的、包括其前后若干词的一个“观测窗口”(w_{-n}，$w_{-(n-1)}$，…，w_0，…，w_{n-1}，w_n) 。本文采用的特征模板如表 2 所示。

表 2 藏文分词模型的特征模板

特征	含义
w_{-2}	中心词前面的第二个词
w_{-1}	中心词前面的第一个词
w_0	中心词
w_1	中心词后面的第一个词
w_2	中心词后面的第二个词

3.2.4 未登录词的处理

虽然我们的训练语料足够大了，但是对人名、地名、组织机构名等的命名实体的覆盖面有限，不可避免地会遇到一些在训练语料中没有出现的词，在这里把这类词称之为未登录词。未登录词的正确标注是分词的一个难点，其标注结果的好坏，会直接影响到整个分词的正确率。解决未登录词正确标注的方法有两种，第一种是在训练语料中覆盖足够多的人名、地名、组织机构名；第二种方法是通过总结规则来提高未登录词标注的准确率。本文基于以上两种思想：在人民网藏语频道 2014 年的全年的共 6000 多篇藏文文章中提取了 14077 条人名，5359 条地名，6899 条组织机构名，共 26335 条命名实体加入训练集中，同时整理了藏语常用地名、人名、组织机构名实体库。

3.3 基于知识融合的藏文分词

我们通过总结 CRFs 分词结果的错误，并对错误进行分析，归纳总结了基于藏文自身知识的分词规则，并通过这些知识对 CRFs 的结果进行校正。主要针对非藏文字符切分错误，藏文黏着词识别错误，停用词切分错误，一些典型的人名、地名、组织机构名的识别错误等问题分别总结了规则。首先列举几个基于 CRFs 的识别结果中的典型例句和经过模型识别后的错误标注序列以及其正确的标注序列，然后针对这些典型例句错误标注序列进行分析。

3.3.1 非藏文字符的识别错误修正

CRFs 切分结果：　ཀུང་གོ/གཞར་ཐ/དབུ་བརྒྱ/ནས/ཚ/65/འགོར་བ/ར/རྟེན་འབྲེལ/ཞུ/།

正确的切分结果：　ཀུང་གོ/གཞར་ཐ/དབུ་བརྒྱ/ནས/ཚ/65/འགོར་བ/ར/རྟེན་འབྲེལ/ཞུ/།

切分错误：　ཚ65 应切分 ཚ/65

导致该类错误的原因是有两种，一是由于语料中存在一定量的非藏文字符，而本文所采用的基于 CRFs 的方法是对藏文音节序列的标注，我们将未分词的藏文语料按照音节序列交给 CRFs 模型参数去识别时，会存在藏文字符和非藏文字符组合当成一个音节，这样训练集中不存在这样的音节而导致错误。二是训练集中本身就存在藏文字符和非藏文字符的组合当成一个音节的现象而导致分词错误。

针对该类错误我们定义如下规则：设 S 表示待切分的藏文句子，$S=\{w_0,w_1.....w_i.....,w_n\}$，$(0<i<n)$，$w_i$ 表示每一个音节。用 U 表示非藏文字符集合，$U=\{D,E,C,P\}$，用 $u_j(0<j<n)$ 表示非藏文字符集合 U 中的元素，其中 D 是时间和数字的集合，例如：“123”，“3.14”，“30%”等，E、C 分别表示英文和汉文字符，P 表示标点符号，包括中英文标点符号、半全角标点符号。

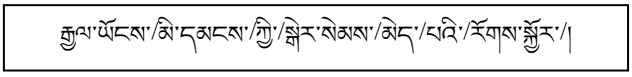
规则 1：

如果 $w_i \in U\,(i \neq 0)$；则将 w_i 单独从集合 S 中切分出来。

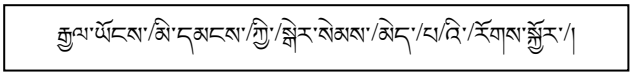
在未分词的语料按照每个音节分开之前先通过该规则将所有非藏文字符单独切分出来，这样避免了交给 CRF 模型参数去识别时，藏文字符和非藏文字符的组合当成一个音节而导致的错误。在得到 CRF 分词的结果之后，再通过该规则处理一次，这样避免了训练集中本身就存在藏文字符和非藏文字符的组合当成一个音节的现象而导致分词错误。

3.3.2 黏着词的识别错误修正

CRF 切分结果：



正确的切分结果：



切分错误： བའི་ 应切成 བ/འི་

导致该类错误的原因是对藏文中黏着词的识别不准确，针对这类错误我们首次引入了词频的信息。首先我们统计了在大规模的训练语料中出现的所有包含黏着词的音节的出现频次，在我们的训练集中总共出现了 101265 条包含黏着词的音节，去重后仅有 305 条不重复的包含黏着词的音节，从中不难发现这些包含黏着词的音节的重复率很高。我们分别计算了每个包含黏着词的音节在训练语料中所占的比例 f_c。 f_c的计算方法如下：

$$f_c = \frac{\text{该音节作为黏着形式出现的频次}}{\text{该音节出现的总频次}} \qquad (6)$$

我们以前十个出现次数最多的包含黏着词的音节作为例子，如表 3 所示：

表 3 前十个出现次数最多的包含黏着词的音节

音节	训练集中出现的总次数	作为黏着形式出现的次数	f_c
བའི་	25808	25301	0.9803549
བར་	10916	6471	0.59279954
བའི་	9032	8978	0.99402124
བར་	4426	2526	0.75071847
བས་	3702	3473	0.9381415
བས་	3560	2763	0.7761194
རྒྱུ་	3349	3346	0.9991042
དོར་	2596	2579	0.9934515
ཚོར་	2329	1736	0.7453843

我们对不同的 f_c值进行了实验，图 2 给出了 f_c对黏着词判断的影响：

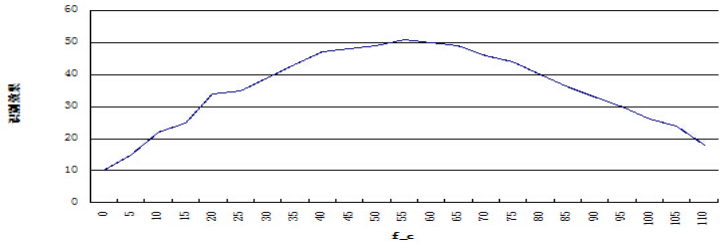


图 2 f_c取值对黏着词识别的效果图

从该表我们不难发现，这几个音节在文本中作为黏着形式的词的概率 f_c 都很高。 我们定义以下规则：将所有符合 $f_c > f$ 的音节加入集合 N，N 表示常用包含黏着词的音节集合，n_j 表示集合 N 中包含的元素。其中 f 是我们自定义的阈值，从实验数据可以得出 f 取值为 55 时效果最佳。

规则 2

如果 $w_i \in \mathrm{N}\ (i \neq 0)$ 且 $f_c > f$， 则将 w_i 判断为带有黏着形式的藏文音节。

3.3.3 停用词的分词错误修正：

由于目前还没有学界公认的藏文停用词表，本文所指的停用词包括如下内容：

无歧义的藏文格助词：例如，“གི་”，“ཀྱི་”，“གྱིས་”，“ཀྱིས་” 等；

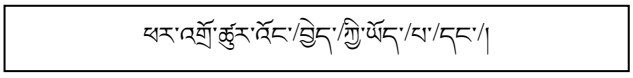
一些功能词：例如，“དང་”，“དེ་”，“རེད་”，“ཡོད་”，“ཡིན་” 等；

藏文标点符号：例如，“།།”，“॥”，“།” 等；

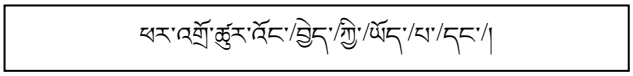
藏文数字词：例如，“གསུམ་”，“༧༣”，“ཞེ་གཅིག་”，“ས་ཡ་” 等。

藏文时间词：例如，“ས་ག་ཟླ་བ་”，“རབ་བྱུང་བརྒྱད་པ་”，“ཞགས་རྟ་ལོ་” 等。

CRFs 切分结果：



正确的切分结果：



切分错误： ཀྱི་ཡོད་ 应切成 ཀྱི་/ཡོད་

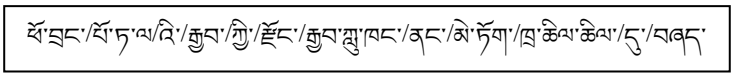
该类错误是本该分开的格助词在 CRF 识别结果中没能分开。例如“རྒྱུ་རེད་”、“ཀྱི་འདུག་” 等。针对这类错误我们整理了藏语常用停用词表。对于这类词我们定义如下规则：设 Sw (stop words)表示停用词集合。

规则 3

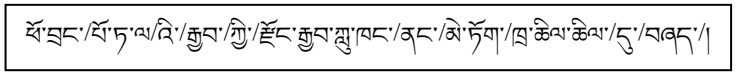
如果： $w_i \in SW \quad (i \neq 0)$， 则将 w_i 单独从集合 S 中分出来。

3.3.4 未登录词识别错误修正

CRF 切分结果：



正确的切分结果：



切分错误： རྩོང་/རྒྱབ་ཀྱུ་ཁང་ 应切成 རྩོང་/རྒྱབ་ཀྱུ་ཁང་

导致该类错误的原因是因为训练集中不包含这个命名实体，从而 CRF 未能识别出来。针对该类错误，我们整理了藏语常用人名、地名、组织机构名实体库。对于这类词我们定义

如下规则：设 T 表示常用实体库，t_j 表示集合 T 中的元素。

规则 4

如果：$w_i \in T \ (i \neq 0)$，则将 w_i 单独从集合 S 中分出来。

4 实验结果

4.1 实验数据

虽然藏文信息处理已进行了多年的研究,但至今没有公开的语料库，因此本实验的训练集语料数据来源是西藏新闻网、人民网藏语频道、青海藏语广播网和新华网等主流媒体的藏语网站。所涉及的领域范围包括新闻、娱乐、诗歌、文化、宗教不同类别的文章。具体的实验数据如下表 4 所示：

表 4 实验数据详细情况

类别	大小	句子数（万）	音节数（万）	词数（万）
总共	35.1M	16.51	217.3	109.74
训练集	31.5M	14.85	195.57	98.77
封闭测试集	3.5M	1.63	21.51	10.86
开放测试集	3.6M	1.66	21.73	10.97

4.2 实验平台

本文实验都是在 PC 机环境下完成的，操作系统是 Win7，使用条件随机场模型进行训练和测试，采用的是 CRF++0.58。CRF++是一个实现了条件随机场模型的工具，大量应用于序列数据的标注和分割，具有良好的通用性，现在已经广泛运用于自然语言处理各个领域的研究和应用中，比如分词、词性标注、命名实体识别、信息抽取等。

4.3 评测指标

我们用 R、P、F 分别表示召回率、正确率、F 值。则 R、P、F 的计算方法公式如下所示：

$$R = \frac{\text{正确的切分的词数}}{\text{文本总词数}} \times 100\% \qquad (7)$$

$$P = \frac{\text{正 确切分词数}}{\text{切分总次数}} \times 100\% \qquad (8)$$

$$F = \frac{2 \times R \times P}{R + P} \times 100\% \qquad (9)$$

4.4 藏语分词结果

我们分别对仅使用 CRF 模型的分词结果和使用规则校正后的分词结果做了比较，如表 5 所示：

表 5 CRF 和规则相结合分词结果

测试类型	结果比较	R	P	F
封闭测试	CRF	0.9849	0.9856	0.9852
	CRF+规则	0.9991	0.9990	0.9990
开发测试	CRF	0.9471	0.9373	0.9421
	CRF+规则	0.9611	0.9603	0.9606

从上表我们可以看出加入本文总结的规则对基于条件随机场模型的藏文分词进行校正之后比起仅使用 CRF 模型在分词的 R、P、F 都有了明显提高。主要是对非藏文字符的切分和黏

着词的再识别以及停用词的再切分都对分词的准确率的提升起到了很好的作用。

在开放测试中，采用本文的 CRF 和规则相结合的方法，分词的 R、P、F 等指标值均达到了 96%，说明基于本文的藏语分词方法可以取得较好的分词效果。在封闭测试中，分词的各项指标均超过了 99%，虽然是在实验条件下的分词结果，但是可以说明利用条件随机场和规则相结合的分词方法对于藏语分词有理想的预期效果。

经过与其他学者的藏文分词研究结果比较可以看出，本文提出的条件随机场和规则相结合的分词方法的分词结果在各项指标上均有提升。

4.5 总结

本文在前人研究的基础上根据藏语的特点实现了一种基于 CRF 和规则相结合的藏语分词系统，通过基于字标注的 CRF 模型分词方法和依照藏文独特的语法特点，使用规则对 CRF 分词结果进行校正，取得了很好的分词效果。分析分词错误的结果集发现，大部分错误都集中在未登录词的识别错误上，接下来，我们希望通过加入更多的藏语语法规则来减少分词系统中对于人名、地名、机构名等命名实体的识别错误。

参考文献

[1] 孙茂松,邹嘉彦.汉语自动分词研究评述[J].当代语言学. 2001, 3(1): 22-32.

[2] 罗秉芬,江荻.藏文计算机自动分词的基本规则[C].中国少数民族语言文字现代化文.北京: 民族出版社, 1999

[3] 扎西次仁.一个人机互助的藏语分词和词登录系统的设计[C].中国少数民族语言文字现代化文集.北京:民族出版社, 1999.

[4] 陈玉忠,李保利,俞士汶等.基于格助词和接续特征的藏文自动分词方案[J].语言文字应用 .2003,(01):75-82.

[5] 陈玉忠,李保利,俞士汶.藏文自动分词系统的设计与实现[J].中文信息学报.2003, (03):15-20.

[6] 才智杰.班智达藏文自动分词系统的设计与实现[J].青海师范大学民族师范学院学报.2010(002):75-77.

[7] Norbu S, Choejey P, Dendup T, et al. Dzongkha word segmentation[C]. Proceedings of the 8th Workshop on Asian Language Resources. 2010. 95-102.

[8] 史晓东,卢亚军. 央金藏文分词系统[J]. 中文信息学报. 2011, 25(4): 54-56.

[9] Liu Huidan Nuo Minghua et al.Tibetan Word Segmentation as Syllable Tagging Using Conditional Random Field[C]. PACLIC.2011.168-177.

[10] 洪铭材，张阔，唐杰，等.基于条件随机场(CRFs)的中文词性标注方法[J].算机科学，2006，33(10):146-151.

[11] 魏欧，孙玉芳.基于非监督训练的汉语词性标注的实验与分析[J].计算机研究与发展，2000，37(4):477 － 482

[12] 李亚超，加羊吉，宗成庆等.基于条件随机场的藏语自动分词方法研究与实现[J].中文信息学报，2013.7

[13] 康才畯，藏语分词与词性标注研究[C].上海师范大学.2014.5

作者简介：洛桑嘎登（1989—），男，硕士，自然语言处理。Email:gaden168@163.com；杨媛媛（1986—），女，博士，计算语言学。 Email:493117207@qq.com；赵小兵（1967—），女，教授，计算语言学。Email:nmzxb_cn@163.com。

（注：一寸照片见下一页）

 洛桑嘎登  杨媛媛  赵小兵