文章编号：1003-0077（2011）00-0000-00

TIP-LAS:一个开源的藏文分词词性标注系统 *

李亚超，江静，加羊吉，于洪志

（甘肃省民族语言智能处理重点实验室（西北民族大学），甘肃 兰州 730030）

摘　要：TIP-LAS 是一个开源的藏文分词词性标注系统，提供藏文分词、词性标注功能。该系统基于条件随机场模型实现基于音节标注的藏文分词系统，采用最大熵模型，并融合音节特征，实现藏文词性标注系统。经过试验及对比分析，TIP-LAS 藏文分词系统词性标注系统取得了较好的实验效果，系统的源代码可以从网上获取。希望该研究可以推动藏文分词、词性标注等基础工作的发展，提供一个可以比较、共享的研究平台。

关键词：藏文；分词；词性标注；条件随机场；最大熵

中图分类号：TP391　　**文献标识码：**A

TIP-LAS: An Open Source Toolkit for Tibetan Word Segmentation and Part of Speech Tagging

LI Yachao, JIANG Jing, JIA Yangji, YU Hongzhi

(Key Laboratory of National Language Intelligent Processing (Northwest University for Nationalities), Lanzhou, Gansu 730030, China)

Abstract: TIP-LAS is an open source toolkit for Tibetan segmentation and POS tagging, which has the function of Tibetan word segmentation and POS tagging. The toolkit implements the Tibetan segmentation system based on syllable tagging with the conditional random fields model, and integrates the maximum entropy model with syllables features to implement the Tibetan POS tagging system. By the experiment and comparative analysis, The System have achieved good experimental results and the source code can be obtained from the Internet, the experimental corpus have also been shared in the Internet. We are expected that the study can promote the development of Tibetan word segmentation, POS tagging and other infrastructure work, and can provide a research platform for comparison and sharing.

Key words: Tibetan; word segmentation; part of speech tagging; conditional random fields; maximum entropy

1　引言

藏文是一种拼音文字，有 30 个辅音字母和 4 个元音字母，由这些字母组成音节，由音节构成词，音节之间用音节点“·”（tsheg）作为分隔符。藏语词语之间没有明显的分隔符来进行区分，因此藏语信息处理首先要面对分词问题。虽然藏语是一种古老的语言，但是对于词汇类别的研究较晚，以往大都是针对藏语构词、形态变化进行的研究 [1,2]。从整体上看，

* **收稿日期：**　　**定稿日期：**

基金项目：西北民族大学中央高校基本科研业务费专项资金资助项目（31920140064）。

藏文分词、词性标注研究基础较弱。在藏文分词、词性标注研究上，没有形成一个公认的或者成熟的分词方法，更没有共享的开源系统。

藏文分词研究相对较早，1999 年发表的“一个人机互助的藏文分词和词登录系统的设计”可以看作是藏文分词研究开始的标志 [3]。其后，陈玉忠等 [4]提出了一种基于格助词和连续特征的书面藏文自动分词方法，该方案在处理切分歧义，解决未登录词问题，提高藏文分词的效果上具有很高的实用价值。以上研究是较为典型的基于规则的藏文分词方法。基于统计的藏文分词方法是最近几年兴起的研究，刘汇丹 [5]等采用格助词分块并识别临界词，然后采用最大匹配方法分词，并进行紧缩词识别，并形成了较为有效的分词系统。孙萌 [6]提出了一种基于判别式感知机模型的藏文分词方法，重点研究最小构词粒度和分词结果重排序对藏文分词的影响，该方法在基于音节的分词系统上加入基于词图的重排序模块，在感知机模型上融合了词典信息。李亚超 [7]研究了基于条件随机场的分词方法，重点解决了紧缩词识别问题。基于条件随机场模型，基于判别式感知机模型，以及基于 HMM 模型的分词方法是藏文分词研究的主要方法。这些分词方法大都以统计模型为基础，融合词典或者是藏语语言特征。以上藏文分词研究的源代码和实验语料都没有公开，加上实验语料规模大都较小，并且融合了较多的语言规则，实验结果难以进行有效的对比。

针对信息处理用藏文词性标记研究起始于 2005 年，才藏太 [8]在班智达汉藏公文翻译系统中对提出了藏文词性标记问题。苏俊峰 [9]研究了基于 HMM 的藏文词性标记方法。扎西加 [10]基于藏文中虚词发挥的功能，结合标注语料库实现了藏文自动分词和词性标记一体化处理模型。史晓东等 [11]采用 HMM 方法将汉语分词系统 Segtag 移植到藏语分词中，其中分词准确度为 93%，词性标注准确度为 83.17%，该系统是较早的实用的藏文分词词性标注系统。华却才让 [12]研究了基于感知机训练模型的判别式藏文词性标注方法，并且实现了相应的词性标注系统“TiPosTag”。于洪志等 [13]以最大熵模型为基本框架，根据藏文的构词特征研究了融合语言特征的最大熵藏文词性标注模型。以上研究都是针对藏文某种语言学特征，采用统计模型进行建模，最后实现相应的词性标记方法。以上的研究方法，实验条件和实验语料不统一，实验结果相差较大。

论文的其余部分安排如下：第二部分阐述 TIP-LAS 藏文分词词性标注系统的基本思路和系统特点；第三部分介绍基于条件随机场的藏文分词方法；第四部分介绍基于最大熵的藏文词性标注方法；第五部分给出实验数据，并进行结果分析；最后第六部分为全文总结和展望。

2. 本系统的特点

由于藏文分词、词性标注语料难以获得，更没有形成规模的共享语料可以使用，已有的研究都是在私有语料上取得的实验结果，实验语料规模大都较小，实验结果相差很大，难以进行有效的对比。本系统希望解决限制成熟的自然语言处理方法在藏文上应用的关键问题，尽量减少对藏语语言知识库的依赖，尽可能提高藏文分词、词性标注方法的可移植性。为此，延续了前期关于藏文分词、词性标注的研究成果，开发了较为成熟的，以可实用为目的藏文分词、词性标注系统，命名为“TIP-LAS”，并在规模较大的语料上进行了实验。分词、词性标注系统源代码可以从以下地址获得 [1]。

TIP-LAS 集成藏文分词、词性标注功能，该系统由 C++实现，提供跨 Linux，Windows 平台功能，分为藏文分词系统，词性标注系统两大模块。藏文分词系统基于条件随机场模型，实现了基于音节标注的藏文分词方法，藏文词性标注系统基于最大熵模型，并融合了音节特

[1] https://github.com/liyc7711/tip-las

征。该系统的准确度和速度已经基本满足实际应用要求。

3 基于条件随机场的藏文分词方法

3.1 条件随机场模型

条件随机场(Conditional Random Field, CRF)是 Lafferty 等提出的一种统计的序列标记模型 [14]。在本文分词系统中，把藏文分词看成是序列标记问题。在序列标记问题中生成一个基于无向图（undirected graph）$G = (V,E)$的一阶线性链式 CRF（linear-chain CRF）。V 是随机变量 Y 的集合 $Y = \{Y_i | 1 \le i \le n\}$，对于输入一个句子的 n 个需要标记单元，$E = \{(Y_{i-1}, Y_i) | 1 \le i \le n\}$是 n-1 个边构成的线性链。对于每个句子 x，定义两个非负因子：

对于每个边：$\exp(\sum_{k=1}^{K} \lambda_k f_k(y_{i-1}, y_i, x))$ (1)

对于每个节点：$\exp(\sum_{k=1}^{K'} \lambda'_k f'_k(y_i, x))$ (2)

f_k是一个二值特征函数，K 和 K'是定义在每个边和相应节点的特征数量。

给定一个需要标记的序列 x，其对应的标记序列 y 的条件概率为：

$$P(y \mid x) = \frac{1}{Z(x)} \exp(\sum_{i,k} \lambda_k f_k(y_{i-1}, y_i, x) + \sum_{i,k} \lambda'_k f'_k(y_i, x)) \quad (3)$$

Z(x)是归一化函数。给定训练集 D，训练模型的参数是用来最大化条件似然值。当给定了要标记的序列 x，其对应的标记序列 y 由参数 $Argmax_{y'}P(y'|x)$给出。

本系统使用 CRF 模型来实现藏文音节标注任务。

3.2 分词特征选择

在分词特征选取上选取音节类别和音节上下文特征信息，音节上下文特征模板如表 1 所示。音节类别，在本文中分为藏语音节、藏语标点符号、汉语标点符号、英文字母等。

表 1 分词特征模板

特征	说明
$C_n(n = -2,-1,0,1,2)$	当前音节的前后第 n 个音节
$C_nC_{n+1}(n = -2,-1,0,1)$	连续的两个音节
$C_{-1}C_1$	当前音节前、后的两个音节

3.3 基于 CRF 的藏语分词

基于字标注的分词方法中，需要对每一个字在词中的位置信息进行标注，根据前期研究结果，在系统中选用“BMES”标记集，根据每个藏文音节在词中出现的位置，给予不同的标签，B 代表词的左边界，E 代表词的右边界，M 代表词的中间部分，S 代表单音节词，标记示例如表 3 所示，超过 3 音节的词中间部分都标记为 M。在分词中，把输入的原始藏文文本切分成音节序列，音节序列包含藏文音节，英文，汉语标点符号等，采用 CRF 模型对音节进行位置标注，根据标注结果还原出分词结果。

表 2 音节标记示例

音节数	藏语词汇	标记示例
1	ང་(我，nga)	ང་/S
2	སློབ་མ(学生，slob ma)	སློབ་/B མ/E
3	གསར་འགོད་པ་(记者，gsar vgod pa)	གསར་/B འགོད་/M པ་/E
4	རྒྱུན་ལས་སྒྲུབ་ཞི(常务主席，rgyun las kruvu zhi)	རྒྱུན་/B ལས་/M སྒྲུ་/M ཞི/E

4 基于最大熵的藏文词性标注方法

4.1 最大熵模型

最大熵模型能够融合复杂的特征，在英语、汉语等语言词性标注研究中取得了较好的效果。该模型最初由 E.T.Jaynes 在 1950 年提出，Della Pietra 等[15]将其应用于自然语言处理中。最大熵原理的基本思想是，首先利用给定的训练样本，选择一个与训练样本一致的概率分布，它必须要满足所有已知的事实。在没有更多的约束和假设的情况下，对于那些不确定的部分，则会赋予均匀的概率分布。熵是用来表示随机变量的不确定性，不确定性越大，熵越大，分布越均匀。最大熵模型：

$$P^* = \arg\max_{p \in C} H(P) \tag{4}$$

$H(P)$是模型P的熵,C是满足条件约束的模型集合，下面需要寻求P^*，P^*的形式如下：

$$P^*(y|x) = \frac{1}{Z(x)} \exp(\sum_i \lambda_i f_i(x,y)) \tag{5}$$

$Z(x)$是归一化常数，表示形式如下：

$$Z(x) = \sum_y \exp(\sum_i \lambda_i f_i(x,y)) \tag{6}$$

λ_i为特征的权重参数。

综合考虑藏文词性标注速度与准确度，本系统选用最大熵模型作为序列标注工具。

4.2 特征选择

4.2.1 上下文特征

一个词的词性很大程度上由其上下文的环境决定，因此当前词的前后 n 个词可以作为判断当前词词性的依据。特征模板如表 3 所示：

表 3 上下文特征模板

特征	说明
$C_n(n = -2,-1,0,1,2)$	当前词的前后第 n 个词
$C_nC_{n+1}(n = -2,-1,0,1)$	连续的两个词
$C_{-1}C_1$	当前词前、后的两个词

4.2.2 词内部特征

藏文属于拼音文字，是形态较为丰富的语言，动词的现在、将来、过去时和命令式是通过词缀及附加词缀来表现的。一般来说藏文动词的屈折形态可以分为同根类型和异根类型两

种。对于词内部特征函数定义为：

$$f(x_i, y_i) = \begin{cases} 1 & f(\text{prefix}(x_i) = \text{“rko” and } y_i = \text{v}) \\ 0 & \text{Otherwise} \end{cases}$$

词汇词尾音节特征函数定义为：

$$f(x_i, y_i) = \begin{cases} 1 & f(\text{suffix}(x_i) = \text{"cag" and } y_i = \text{v}) \\ 0 & \text{Otherwise} \end{cases}$$

词内部信息特征模板如表 4 所示。

表 4 词内部特征

模板	说明
W	当前词
prefix(w)	当前词的词首音节
suffix(w)	当前词的词尾音节

将当前词的词首音节、词尾音节，前、后词，前驱词的词尾音节、后继词的词首音节等特征结合在一起，定义音节特征见表 5 所示。

表 5 音节特征模板

特征	说明
w_0(prefix(w)), w_0(suffix(w))	当前词的首、尾音节
w_{-1}(suffix(w))	前驱词的词尾音节
w_1(prefix(w_1))	后继词的词首音节

5 实验设置与系统对比

5.1 实验准备

分词系统采用的语料为第七届全国机器翻译研讨会(CWMT2011），藏汉报刊政论文献平行语料中的藏语语料部分，共 128 万词。把整体语料按照 3:7 的比例分为测试语料和训练语料。

词性标注系统采用的语料从主流的藏语新闻网站抓取网页正文，语料主要来源是中国西藏网、青海藏语广播网、人民网藏语版等，选取政治、经济、新闻、社会、法律等领域的文本。对获取的藏文生语料，经过分词、词性标注工具处理后，再由人工校对获得词性标注语料，语料统计如表 6 所示。训练语料来源于中国西藏网、青海藏语广播网，共 212 万词，测试语料来源于人民网藏语版，共 46 万词。

表 6 语料统计

网站	网址	词数（万）
中国西藏网	http://tb.tibet.cn/	140
青海藏语广播网	http://www.qhtb.cn/	72
人民网（藏语版）	http://tibet.people.com.cn/	46

5.2 系统性能

表 7 实验结果

系统	R(%)	P(%)	F(%)
分词	95.35	95.32	95.33
词性标注	-	93.90	-

本文分词系统在测试语料上的 F 值达到 95.33%，词性标注准确度达到 93.90%，词性标注系统由于输入的是分好词的序列，所以只计算准确度。

5.3 系统对比

表 8 分词标注系统对比

系统	分词			词性标注
	R(%)	P(%)	F(%)	P（%）
SegT	96.91	96.98	96.95	-
孙萌	96.81	95.70	96.25	-
华却才让	95.19	96.84	95.84	98.26
央金系统	92.47	92.12	92.30	83.17
康才畯	90.85	91.27	91.06	84.30
TIP-LAS	**95.35**	**95.32**	**95.33**	**93.90**

已有公开的藏文分词、词性标注系统采用的方法、语料、词典各异，进行严格的对比较为困难，以下列出主要的几个公开发表的系统实现，希望可以进行近似的结果比较。以下对语料规模的表述依据原文的数据表述方式。

SegT 分词系统采用 3000 句训练语料，1000 句测试语料。该系统使用格助词分块和最大匹配方法进行分词，采用双向切分检测分词歧义并使用预先统计的词频信息进行消歧。最大匹配分词方法对词典的依赖性非常大，需要高质量的分词词典才能实现。该系统为规则和统计相结合的藏文分词系统。

孙萌的系统采用 12942 句语料，共 110K 词语，从中随机选择 500 句作为测试集，剩余的作为训练集。该系统采用感知机模型，在基于音节的分词系统上加入基于词图的重排序模块，采用了分词切分语料和词典等语料资源。

华却才让的系统采用 2.2 万多句词性标注句子为感知机模型训练语料，测试语料 573 句，词性词典是从训练语料、班智达词性词典中获得的 9.3 万多条词语，1.9 千条地名词语，1.6 万条人名词典以及计算机等专用词典中抽取，总共抽取到 12.36 万余条藏语词条。系统在人工建立的 573 句藏语词性标注测试集上，分别做了标准测试和分词标注一体化测试。采用了分词切分语料和词典等语料资源。

央金藏文分词系统移植于基于 HMM 的汉语分词系统 Segtag，采用 2.7M 训练语料（UTF16 编码），以及词典，测试语料 25K。分词、词性标注系统加入了 20 多万条的藏汉人名对照词典，通过构造词图来提高基线系统的效果。

康才畯 [16]的分词系统训练语料约 100 万字，测试语料约 2 万字。在词性标注系统中，训练语料 20 万词，测试语料 320 个句子，采用的标记集有 20 个一级类，52 个二级类。分词系统采用基于条件随机场模型方法，词性标注系统采用最大熵模型。实验中没有加入规则和词典等额外信息。

TIP-LAS 系统在分词、词性标注任务上相对来说效果较好，在分词任务上低于前 3 个系

统，在词性标注上低于第 3 个系统。但是，本文系统没有采用词频，语言规则、词性词典、人名词典等辅助资源。本文系统实验结果全部采用训练语料所包含的特征，目的是为提供一个可以比较的实验平台。

6 总结与展望

本文延续了前期关于藏文分词、词性标注的研究，并对前期研究进行整合，实现相应的软件平台，本系统在分词、词性标注上与已有的系统相比取得了较好的效果，并在本单位的机器翻译、语音翻译等系统中得到实际应用。

开源和共享是自然语言处理研究的发展趋势，英文的开源系统较为丰富，汉语成熟的开源系统有哈尔滨工业大学的 LTP，东北大学的 Niutrans，复旦大学的 FudanNLP 等，这些都是优秀的中文信息处理平台，在推动中文信息处理进步中起着不可替代的作用。藏语信息处理研究基础较弱，没有开源软件可以使用，共享语料也很少。本文把最新的研究成果形成实用的软件系统，把藏文分词、词性标注集成到一个平台里，并公开源代码。希望该研究得到更多人的加入，形成共享的开源平台，推动 TIP-LAS 的不断完善，促进藏文信息处理的发展。

参考文献

[1] 宋金兰．汉藏语形态变体的分化．民族语文，2002,1:29-33.

[2] 龙从军．藏语形容词性语素研究[J]．Journal of Chinese Language and Computing. 2006, 15 (4):193-201.

[3] 扎西次仁．一个人机互助的藏文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，1999：322-327.

[4] 陈玉忠，俞士汶．藏文信息处理技术的研究现状与展望[J]．中国藏学，2003, 04:97-107.

[5] 刘汇丹，诺明花，赵维纳等．SegT：一个实用的藏文分词系统[J]．中文信息学报, 2009, 1(26):97-103.

[6] 孙萌，华却才让，才智杰等．基于判别式分类和重排序技术的藏文分词[J]．中文信息学报，2014, 28(2):61-65.

[7] 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，2013,27(4):52-58.

[8] 才藏太，华关加．班智达汉藏公文翻译系统中基于二分法的句法分析方法研究中[J]．中文信息学报，2005,6(19)7:12.

[9] 苏俊峰．基于 HMM 的藏语语料库词性自动标注研究[D]．西北民族大学，硕士学位论文，2010.

[10] 扎西加，高定国．藏文文本分词赋码一体化研究[J]．西藏大学学报（自然科学版）2012,1(27):57-61.

[11] 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.

[12] 华却才让，刘群，赵海兴．判别式藏语文本词性标注研究[J]．中文信息学报, 2014, 2(28):56-60.

[13] 于洪志，李亚超，汪昆，冷本扎西．融合音节特征的最大熵藏文词性标注研究[J]．中文信息学报，2013,5(27)160:165.

[14] J. Lafferty, A. McCallum and F. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data[C]. In Proceedings of ICML-2001, 2001:282-289.

[15] Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra.A Maximum Entropy Approach to Natural Language Processing[J]. Computational Linguistics, 1996, 1(22):39-71.

[16] 康才畯. 康才藏语分词与词性标注研究[D].上海师范大学，博士学位论文,2014.

作者简介：李亚超（1986—），男，讲师，主要研究领域为词法分析、机器翻译、少数民族语言文字信息处理。Email: liyc7711@gmail.com；江静（1988—），女，助理馆员，主要研究领域为复杂网络。Email: 506775848@qq.com；加羊吉（1985—），藏族，女，博士，副教授，主要研究领域为藏文信息处理。Email: 236164976@qq.com。

作者照片：

作者一 李亚超	作者二 江静	作者三加羊吉