

文章编号: 1003-0077(2019)07-0075-06

## 基于混合策略的藏文虚词识别方法

拉玛扎西<sup>1,2,3</sup>, 才智杰<sup>1,2,3</sup>, 班玛宝<sup>1,2,3</sup>

- (1. 青海师范大学 计算机学院, 青海 西宁 810016;
2. 青海省藏文信息处理与机器翻译重点实验室, 青海 西宁 810008;
3. 藏文信息处理教育部重点实验室, 青海 西宁 810008)

**摘要:** 藏文虚词在歧义消解、句法、句型和语义处理等方面起着重要的语法作用。该文在分析传统藏文虚词研究成果的基础上,统计了面向自然语言处理的藏文虚词及特征,提出了基于规则和最大熵模型相结合的藏文虚词识别策略。实验表明,该方法识别藏文虚词的准确率、召回率和  $F_1$  值分别达 98.39%、98.75%、98.57%。

**关键词:** 自然语言处理;藏文虚词;基于规则;最大熵模型

**中图分类号:** TP391

**文献标识码:** A

## Tibetan Function Word Recognition Method Based on Hybrid Strategy

LAMA Zhaxi<sup>1,2,3</sup>, CAI Zhijie<sup>1,2,3</sup>, BAN Mabao<sup>1,2,3</sup>

- (1. College of Computer Science and Technology, Qinghai Normal University, Xining, Qinghai 810016, China;
2. Qinghai Provincial Key Laboratory of Tibetan Information Processing and Machine Translation, Xinning, Qinghai 810008, China;
3. Key Laboratory of Tibetan Information Processing, Ministry of Education, Xinning, Qinghai 810008, China)

**Abstract:** Tibetan function word plays an important role in ambiguity resolution in both syntax and semantics in Tibetan language. This paper examines the Tibetan function words related to natural language processing, and proposes a Tibetan function word recognition combining rules and Maximum Entropy Model. Experiments show that the accuracy, recall and  $F_1$  value of the proposed method reaches 98.39%, 98.75%, and 98.57%, respectively.

**Keywords:** NLP; Tibetan function words; rule-based; maximum entropy model

## 0 引言

藏文是一种典型逻辑格语法体系的拼音文字<sup>[1]</sup>,由实词和虚词按一定的语法结构组合而成。实词具有具体词汇意义,包括名词、代词、动词、形容词、数词等,可以单独使用;而虚词没有实际意义,包括语法虚词<sup>[2]</sup>(格助词和接续助词)和关联词等,不能单独使用。计算机自动识别虚词对文

本的歧义消解、句法分析、句型及语义处理等具有重要作用,并在藏文分词<sup>[3]</sup>和停用词选取<sup>[4]</sup>等方面有重要的应用价值。现有文献中未见详细面向自然语言处理的藏文虚词特征及其个数的分析,并且没有研究多音节虚词的识别。本文在分析传统藏文虚词研究成果的基础上,初步统计了面向自然语言处理的藏文虚词,并分析了藏文虚词的特征,从而提出了基于规则和最大熵模型相结合的藏文虚词识别策略。

**收稿日期:** 2018-08-06 **定稿日期:** 2018-09-28

**基金项目:** 国家自然科学基金(61866032, 61163018, 61262051, 61662061);国家社会科学基金(13BYY141, 16BYY167, 15BYY167);教育部“春晖计划”合作科研项目(Z2012093, Z2016077);青海省基础研究项目(2017-ZJ-767, 2019-SF-129, 2015-SF-520);“长江学者和创新团队发展计划”创新团队资助项目(IRT1068);青海省重点实验室项目(2013-Z-Y17, 2014-Z-Y32, 2015-Z-Y03);藏文信息处理与机器翻译重点实验室项目(2013-Y-17);青海师范大学 2017、2018 年度创新训练项目

本文组织结构如下:第1节分析藏文虚词识别的研究现状和主要技术方法;第2节归纳并总结传统藏文文法和面向自然语言处理的藏文虚词,确定面向自然语言处理的虚词数量及特征;第3节设计基于规则和最大熵模型相结合的藏文虚词识别方法;第4节实验验证算法的有效性,并对存在的问题进行分析;第5节是结论与展望。

## 1 研究现状

分词既是藏语自然语言处理的一项基础性研究工作,也是一个存在很多难点的研究范畴。陈玉忠等<sup>[2]</sup>在分析藏文文本自动切分难点时指出,藏文分词中较难解决的问题有四类:①由实词—实词、实词—虚词、虚词—实词、虚词—虚词的交集性字段引起的错误;②由实词—实词、实词—虚词、虚词—实词、虚词—虚词的组合型歧义字段引起的错误;③由紧缩词识别引起的错误;④由未登录词引起的错误。在这四类错误中,前三项与虚词的识别有关。因此,藏文虚词(包括紧缩词)的识别问题引起学者们的关注。其中,紧缩词是一种特殊的虚词,学者们先后研究了紧缩词的识别问题。才智杰<sup>[5]</sup>首次提出了紧缩词的“添加—还原法”识别方法,识别准确率达99.83%,取得了理想效果。完么扎西等<sup>[6]</sup>在“添加—还原法”的基础上利用藏文文法规则识别紧缩词,其识别准确率达99.95%。李亚超等<sup>[7]</sup>为解决无法识别未登录词后的紧缩词问题,提出了基于条件随机场的紧缩词识别方法,其识别准确率达98.91%,克服了“还原法”中不能识别“未登录词+紧缩词”的问题。华却才让等<sup>[8]</sup>利用藏文紧缩词识别音节的方法,识别准确率达到99.91%。康才峻等<sup>[9]</sup>采用基于词位的统计分析方法识别藏文紧缩词的准确率为95.89%,解决了未登录词对识别效果的影响。拉玛扎西等<sup>[10]</sup>通过剖析现有藏文紧缩词识别方法,分析藏文字词的特征,有针对性地提出了基于规则、添加—还原法与最大熵模型相结合的藏文紧缩格识别方法,其识别准确率达到99.26%,相比现有准确率,有明显的提高。同样,在一般虚词识别方面,学者们也提出了若干识别方法。赵栋材<sup>[11]</sup>通过建立虚词兼类词典库,在采用正向最大匹配算法对文本切分后,利用不自由虚词的接续规则识别虚词(单音节虚词)。高定国等<sup>[12]</sup>提出了

基于规则的藏文虚词识别方法,其识别准确率达97.08%。拉巴顿珠等<sup>[13]</sup>通过建立虚词兼类词典、单音节词典、规则的不自由虚词词典库等识别藏文虚词。由以上文献可见,特殊虚词紧缩词的识别问题利用统计与规则相结合的方法可以得到解决,但一般虚词的识别还不能满足实际需求。一般虚词的识别主要有两个不足点:①识别方法只用了规则法。由于虚词的多样性,仅依靠规则不能识别出好的效果。正如文献<sup>[12]</sup>在实验分析中指出,在规则法的基础上引入统计方法,可以提高藏文虚词识别率。②没有具体分析虚词的特征,只是笼统地将藏文文法中提到的虚词认定为面向自然语言处理的虚词对象,其识别对象没有完全囊括藏文文本中的虚词。

## 2 藏文虚词及其特征

在藏文虚词识别研究的文献中,没有明确藏文虚词及其数量,因而在自然语言处理的各项研究中没能获得理想的成果。研究面向藏语自然语言处理的藏文虚词识别方法,依据藏文文法理论,并将其具体化,才能取得好的效果。本节通过分析传统藏文文法中虚词的定义及数量,确定了面向自然语言处理的藏文虚词,并分析其特点。

### 2.1 传统藏文文法中的虚词

藏文文法《三十颂》是一部最早阐述藏文文法的专著,里面有专门阐述藏文虚词的内容。《三十颂》从语法功能角度给出了虚词的定性描述:虚词是指按语境添接在实词的前或中或尾部后,使各零散的实词具有一定意义的功能词<sup>[14]</sup>。《三十颂》中罗列的虚词都是单音节虚词。在后续的研究中,学者们对《三十颂》做了很多不同的解读,将虚词按音节数分为单音节虚词和多音节虚词<sup>[15]</sup>。文献<sup>[14,16-18]</sup>解读《三十颂》中对虚词的阐述,罗列了藏文虚词(下文中把这类虚词称为语法虚词),各文献收录的藏文语法虚词数量统计见表1。

传统藏文文法认为表1中的所有语法虚词全都是单音节虚词。事实上,虚词“ཞིང་”等中的“ཅེན་ཞིང་ཞིང་ཞིང་”三个是双音节虚词。在解读《三十颂》中语法虚词的基础上,学者们纷纷讨论了单音节和多音节虚词补遗问题(下文中把这类虚词称为补遗虚词)。藏文补遗虚词统计见表2。

表 1 藏文语法虚词数量统计表

虚词名	文献[14]	文献[16]	文献[17]	文献[18]	虚词名	文献[14]	文献[16]	文献[17]	文献[18]
属格助词	5	5	5	5	语气助词	1	1	1	1
作格助词	5	5	5	5	连词	1	1	1	1
位格助词	7	7	7	7	指示代词	1	2	1	2
从格助词	2	2	2	2	疑问代词	6	6	4	5
呼格	4	—	4	4	否定词	4	2	4	4
待述词	3	3	3	3	指人后缀	9	9	9	9
离合词	11	11	11	11	时态词	3	—	—	4
终结词	11	11	11	11	格助词	23	19	23	23
虚词ཞིང等	15	14	—	14	接续虚词	68	60	48	68
饰集词	3	3	3	3	总数	91	82	71	91

表 2 藏文补遗虚词统计表

文献	单音节虚词数	多音节虚词数	总数
文献[14]	54	303	357
文献[16]	—	65	65
文献[18]	35	29	64
文献[19]	—	137	137

表 2 中的补遗虚词不包含语法虚词,语法虚词在藏文真实文本中经常出现,起到转折、关联等作用。

2.2 面向自然语言处理的藏文虚词

由于自然语言处理的特殊需求,面向自然语言处理的虚词不能直接选用传统藏文文法中规定的虚词,需要分析语法虚词中单音节虚词的语法作用以及在文本中的词性,并对个别在藏文文法中提到的补遗虚词进行相应处理后,才能最终确定虚词识别任务的处理对象。

本文在选取和识别面向自然语言处理的虚词时,遵循以下 5 条原则。

**原则 1** 词缀“བཤོ་བཤོ་མཚོ་ཅན་ལྷན་མཁན་”等 9 个不列入本文识别的虚词。词缀“བཤོ་བཤོ་མཚོ་ཅན་ལྷན་མཁན་”等 9

个语法虚词在语法角度具有虚词的功能,但在实际应用中添加在某一实词的前、中、后成为该实词不可分割的成分。因此这 9 个虚词不应列为面向自然语言处理的单音节虚词。

**原则 2** 否定虚词中的“མིན”(不是)和“མེད”(没有)表示否定意义,具有实际意义且在文本中能独立运用。在面向藏语自然语言处理中将这两个词归为实词。

**原则 3** 藏文多音节补遗虚词中,有很多词具有实际意义,例如,“ང་ཚེ”“པེད་ཅག”“འཕྲུལ”等实词不列为藏文多音节虚词。

**原则 4** 兼类虚词不加区别。例如,“ཅུ”有时为位格助词,有时为疑问代词;“དྲི”有时为待述词,有时为指示代词。本文识别时不区分虚词的类型。

**原则 5** 识别虚词时,虚词按音节数分类,这里的音节数指虚词实际所含音节的个数,与语法虚词中的音节数的意义不同。例如,虚词ཞིང等中的“ཅིན”“ཞིན”“ཤིན”等在语法虚词中被认为是单音节虚词,本文识别虚词时将其归为双音节虚词。

本文从表 1、表 2 罗列的虚词中,遵循以上 5 条原则,确定了面向自然语言处理的 552 个虚词,面向自然语言处理的藏文虚词及其分布如表 3 所示。

表 3 面向自然语言处理的藏文虚词及分布表

面向自然语言处理的虚词 552		虚词类型	总数			比例/%		
		语法虚词	72			13.04		
		补遗虚词	480			86.96		
语法虚词	总数	比例/%	语法虚词	总数	比例/%	全部虚词	总数	比例/%
兼类	50	69.44	自由虚词	12	16.67	单音节	106	19.20
非兼类	22	30.56	不自由虚词	60	83.33	多音节	446	80.80

由表 3 可知,在 552 个面向自然语言处理的藏文虚词中,有 72 个语法虚词和 480 个补遗虚词。72 个语法虚词中兼类虚词有 50 个,480 个补遗虚词中兼类虚词有 16 个。藏文语法虚词中兼类虚词所占比例高达 69.44%,对虚词的识别带来了困难。语法虚词中自由虚词有 12 个,不自由虚词有 60 个,占语法虚词总数的 83.33%,480 个补遗虚词都为自由虚词。从虚词所含音节角度看,单音节虚词有 106 个,多音节虚词有 446 个,可见藏文虚词以多音节为主。

### 2.3 藏文虚词的特征

藏文虚词除了表示语法意义和不能单独使用的共性特征外,还具有以下 5 种个性特征。

#### (1) 黏着特征

虚词中的 la 格助词“ར”、具格助词“ས”、属格助词“འ”、终结词“ང”、饰集词“ང”和离合词“འཕྲུལ”与其前一个音节之间不加音节点组成一个音节。例如,“ངའི་ལྗོན་པོ།”(我的尊严)中的音节“ངའི”为实词“ང”和属格助词“འ”的(的)的紧缩形式。

#### (2) 兼类特征

藏文虚词中兼类虚词所占的比例特别大,主要有两种兼类形式:①虚词与虚词兼类。例如,“གསུང་པོ་ལ་ཕྱི་ཕྱོད་པོ།”(去东方)”和“ཡི་གེ་མི་འབྲི་མཁན་ལྟ་ཅི།”(不写字的是谁?)”两句中的“ལྟ་”同样是虚词,但第一句中的“ལྟ་”为 la 格助词,第二句中的“ལྟ་”为疑问代词。②虚词与实词兼类。例如,“ལོ་ལོ་བཞུགས།”(爬山坡)”中的第一

个“ལ”为实词,意为山坡,第二个“ལ”为 la 格助词。

#### (3) 实词中包含单音节虚词的特征

例如,“བྱི་མེད་རྟ་ཡི་བར་ཁ་ལྷན།”(马不停蹄)”中“བྱི་མེད་(马)”的首个音节与属格助词“ཡི”兼类。

#### (4) 多音节虚词包含单音节虚词的特征

这类虚词主要是关联词,其组词形式可以归为两种:①接续虚词+(虚词|实词)形式。例如,“དེ་ནས་སྐོར་ལ་ཕྱི་ཕྱོད་པོ།”(然后去上学)”中的多音节虚词“དེ་ནས་”中包含单音节接续虚词“དེ”和从格助词“ནས”;虚词“ལྱི་ར་བཏང་”中包含单音节虚词“ལྱི་ར”和实词“བཏང་”。②实词+虚词形式。例如,“ཐེལ་བ་ཆེ་བའི་དབང་གིས་བསྐྱབ་མ་ཐུག།”(忙的没能办事)”中的虚词“དབང་གིས་”为实词+虚词的形式,该虚词中包含单音节具格助词“གིས”。

#### (5) 多音节虚词具有嵌套特征。

藏文多音节虚词中又嵌套另一个多音节虚词。例如,多音节虚词“ཡང་ནས་ཡང་དུ་”中嵌套了两个多音节虚词,分别为“ཡང་ནས་”和“ཡང་དུ་”;多音节虚词“གང་ནས་གང་ལ་”中嵌套了多音节虚词“གང་ནས་”和“གང་ལ་”。

## 3 藏文虚词识别

### 3.1 藏文虚词识别策略

本文采用逆向最大匹配法和最大熵模型相结合的混合策略识别藏文虚词。其识别模型如图 1 所示。

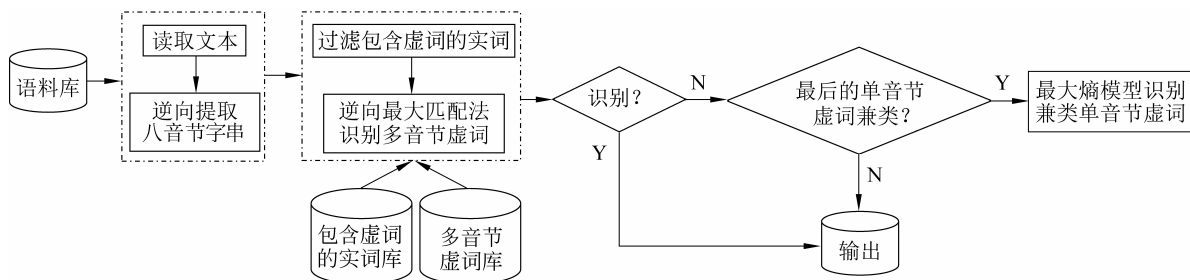


图 1 是根据藏文虚词特征提出的基于规则法和最大熵模型相结合的混合策略模型。针对虚词中具有黏着特征的紧缩词识别已有很多研究,其识别准确率达 99.83% 以上,本文运用了文献[5,10]中提出的“添加—还原法”和基于规则、添加还原法与最大熵模型相结合的藏文紧缩词识别方法,具体参见文献[5,10]。针对藏文虚词的第(4)类特征,文章采用多音节虚词优先识别策略,因此,基于混合策略的

藏文虚词识别模型包含多音节虚词识别模块和单音节虚词识别模块。

多音节虚词识别模块在“包含虚词的实词库”中对文本预处理中逆向提取的 8 音节字串进行查找,若找到,则可断定 8 音节字串中无虚词;否则,在“多音节虚词库”上采用逆向最大匹配法判断是否为多音节虚词。这里只提取 8 音节字串的原因是藏文多音节虚词中最大音节数为 8,而且“包含虚词的实词

库”中的最大音节数也不超过 8 个。其中,“包含虚词的实词库”含 719 个词条,“多音节虚词库”含 446 个词条。

单音节识别模块首先判断多音节模块未能识别的最后一个单音节虚词是否为兼类词,若该单音节虚词不是兼类虚词,则一定为虚词;否则,该单音节有可能是虚词,也有可能是实词。然后,对这个单音节用最大熵模型判别其是否为虚词。由于单音节兼类虚词有 33 个,因而判别虚词的兼类性也比较简单。

3.2 最大熵特征模板

Jaynes 于 1957 年首次提出最大熵原理,被广泛应用于自然语言处理领域。其基本原理是,在已知部分信息的前提下,关于未知分布最合理的推断应该符合已知信息最不确定或最大随机的推断<sup>[20]</sup>。藏文虚词识别可看作是一个序列标注问题,标注时对每个对象随机标注一个标签,并建立已知特征  $x$  的条件下输出标签  $y$  的概率分布模型  $p(p \in P)$ 。其中, $x$  属于上下文信息集  $X(x \in X)$ ,  $y$  属于对应的标签集  $Y(y \in Y)$ 。从训练集中可获得  $N$  个样本集,即  $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ ,根据这些样本可以定义一个事件空间,其特征是一个二值函数  $f: X \times Y \rightarrow \{0, 1\}$ ,其定义如式(1)所示。

$$f(x,y) = \begin{cases} 1 & \text{if } x = x_i \text{ and } y = y_i, (x_i, y_i) \\ 0 & \text{otherwise} \end{cases}$$

则模型  $p$  的熵为:

$$H(p) = - \sum_{x,y} p(x,y) \log[p(x,y)] \tag{1}$$

从式(1)中可得出最大熵模型,如式(2)所示。

$$P^* = \arg \max_{p \in C} H(p) \tag{2}$$

式(2)中的  $C$  为符合约束条件的模型集合,然后计算满足  $C$  条件的最大  $p^*$ ,如式(3)所示。

$$P^*(y|x) = \frac{1}{z(x)} \exp \left[ \sum_i \lambda_i f_i(x,y) \right] \tag{3}$$

其中, $z(x)$ 是归一化常数,并有式(4)。

$$z(x) = \sum_y \exp \left[ \sum_i \lambda_i f_i(x,y) \right] \tag{4}$$

式(3)、式(4)中的  $\lambda_i$  为模型参数,即特征  $f_i$  对应的权重  $\lambda_i$ ,可通过 IIS 算法来估计。

藏文虚词识别时,把原始文本内容进行序列标注,如“ཁོ་མོ་ཁོ་མོ་ཉི་ཤེས་པ་(犁地种田)”的文本语料,经过标注后的训练语料为“ཁོ་Fམོ་Nཁོ་Fཉི་Tཤེས་Fཔ་Nཤེས་Nཔ་N”。其中  $T$  表示虚词, $F$  表示兼类虚词, $N$  表示非虚词。首先

将 $\{T,F,N\}$ 作为标签集,每个音节及其上下文信息作为输入值;然后使用最大似然估计统计语料中每个特征概率;最后选取模型返回的每个标签的最大输出概率。

最大熵模型中,如何针对研究对象选择有效的上下文特征是一个关键问题。本文根据藏文词语音节的分布特点及上下文激发环境确定模型,并抽取特征模板。本文选取的特征模板如表 4 所示。

表 4 特征模板

序号	原子模板	模板意义
1	$F\_word$	当前音节(拟虚词)
2	$L\_F\_word$	当前虚词及其前一个音节
3	$F\_word\_L$	当前虚词及其前两个音节
4	$L_e\_F\_word$	抽取前一音节的后加字母集

4 实验数据及分析

为了验证本文提出的藏文虚词识别方法的有效性,我们从青海师范大学才智杰教授研究小组建立的藏语语料库中选取了含 30 404 个音节的语料作为测试语料,语料领域包括政治、教材、历史、小说、新闻等五种题材。语料中含 9 187 个藏文虚词,利用本文提出的藏文虚词识别方法正确识别出了 9 040 个虚词,共出现 187 个识别错误,实验数据见表 5。

表 5 虚词识别实验数据

评测项	准确率%	召回率%	$F_1$ %
评测值	98.39	98.75	98.57

由实验数据可以看出,本文提出的基于规则和最大熵模型相结合的藏文虚词识别方法的准确率、召回率和  $F_1$  值分别达到了 98.39%、98.75%和 98.57%。通过分析 187 个识别错误的文本,发现主要原因是未能解决多音节虚词兼类的问题。例句“སྤྱད་སྤྱད་དེ་ཡང་།(那枯树很轻)”,因藏文多音节虚词中收录了虚词“དེ་ཡང”,在识别过程中多音节虚词识别模块错误地将“དེ་ཡང”识别为多音节虚词。事实上,这里“སྤྱད་སྤྱད”和“ཡང”为实词,“དེ”为单音节虚词。这类多音节兼类虚词有“སྤྱད་སྤྱད་དེ་ཡང་།ལ་སྤྱད་སྤྱད་དེ་ཡང་།ལ་སྤྱད་སྤྱད་དེ་ཡང་།”等 7 个,如果在多音节虚词识别模块中增加判断多音节虚词兼类可以得到解决。

## 5 结论与展望

藏语虚词识别既是藏语自然语言处理的一项基础性工作,也是一项具有挑战性的研究工作,在藏文分词和停用词选取等方面有重要的应用价值。本文重点探讨了面向自然语言处理的藏语虚词及其语法特征,确定了面向自然语言处理的虚词及数量,提出了规则法和最大熵模型相结合的藏文虚词识别混合策略。实验表明,该方法识别藏文虚词的准确率、召回率和  $F_1$  值分别达 98.39%、98.75%、98.57%。今后在该研究成果的基础上,将进一步研究藏文分词及停用词选取技术,为藏文词向量表示奠定基础。

## 参考文献

- [1] 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.
- [2] 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,2003,17(3): 15-20.
- [3] 陈玉忠,李保利,俞士汶,等.基于格助词和接续特征的书面藏文分词方案[J].语言文字应用,2003(1): 75-82.
- [4] 珠杰.藏文信息处理中若干关键技术研究[D].成都:西南交通大学博士学位论文,2016.
- [5] 才智杰.藏文自动分词系统中紧缩词的识别[J].中文信息学报,2009,23(1): 35-37.
- [6] 完么扎西,尼玛扎西.藏语自动分词中的几个关键问题的研究[J].中文信息学报,2014,28(4): 132-139.
- [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): 51-58.
- [8] 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.
- [9] 康才峻,龙从军,江获.基于词位的藏文黏写形式的切分[J].计算机工程与应用,2014,50(11): 218-222.
- [10] 拉玛扎西,才智杰,扎西吉.藏文紧缩格识别方法[J].计算机应用研究,2019,36(4): 1080-1083.
- [11] 赵栋材.基于虚词切分的藏文分词系统的设计与实现[J].西藏大学学报(自然科学版),2012,27(2): 61-65.
- [12] 高定国,扎西加,赵栋材.计算机识别藏语虚词的方法研究[J].中文信息学报,2014,28(1): 113-117.
- [13] 拉巴顿珠,欧珠,赵栋材.藏文自动分词系统中虚词识别算法研究[J].计算机应用与软件,2017,34(9): 299-301.
- [14] 郭须·扎巴军乃.简述藏文语法中的虚词个数[J].西藏研究(藏文版),1992(2): 15-50.
- [15] 珍贝益西扎巴.语门文法概要[M].北京:民族出版社,1980.
- [16] 吉太加.现代藏文语法通论[M].兰州:甘肃民族出版社,2009.
- [17] 才旦夏茸.藏文文法[M].兰州:甘肃民族出版社,2005.
- [18] 格桑居冕,格桑央京.实用藏文文法教程[M].成都:四川民族出版社,2004.
- [19] 邓戈.藏文语法中的多音节虚词的补遗研究[J].西藏大学学报(藏文版),2015,4: 108-122.
- [20] 宗成庆.统计自然语言处理[M].北京:清华大学出版社,2013.



拉玛扎西(1994—),硕士研究生,主要研究领域为藏文信息处理、藏语自然语言处理。  
E-mail: lhamatashi@outlook.com



班玛宝(1992—),硕士研究生,主要研究领域为藏文信息处理、藏语自然语言处理。  
E-mail: 1402554093@qq.com



才智杰(1970—),通信作者,博士,教授,硕士生导师,主要研究领域为藏文信息处理、藏语自然语言处理。  
E-mail: czjqhsd@163.com