融合无监督特征的藏文分词方法研究

李亚超，加羊吉，江静，何向真，于洪志

（西北民族大学 中国民族语言文字信息技术重点实验室，甘肃 兰州 730030）

摘要：藏文分词是藏文信息处理的基础性关键问题，目前基于统计方法的藏文分词特征大都采用音节位置特征和类别特征等。本文研究基于序列标注的藏文分词系统，并在此基础之上增加了从无标注语料中抽取的无监督特征。实验结果表明，本文从无标注数据中提取出的无监督特征较为有效，并且可以和有监督的分词模型融合到一起，显著提高了基线藏文分词系统的效果。

关键词：藏文；分词；序列标注；无监督；特征

中图分类号：TP391　　**文献标识码：**A

Study on Fusion of Unsupervised Features for Tibetan Word Segmentation

LI Yachao, JIA Yangji, JIANG Jing , HE Xiangzhen, YU Hongzhi

(Key Lab of Chinese National Linguistic Information Technology,
Northwest University for Nationalities, Lanzhou 730030)

Abstract: Tibetan word segmentation (TWS) is an important problem in Tibetan information processing, while the current TWS features are mostly adopt the syllable position and syllable categories. The paper studied the TWS as syllable tagging problem, and added the unsupervised features distracting from unlabeled corpus. The experimental results show that the unsupervised features distracted from unlabeled corpus is fast and effective and can be combined easily with the supervised segmentation model. This significantly increases the effect of the baseline TWS.

Key words: Tibetan; word segmentation; sequence labeling; unsupervised; feature

1　引言

分词（Word Segmentation）属于自然语言处理中的经典问题，是将连续的单位（unit）序列按照一定的规范重新组合成词序列的过程。藏语是一种拼音文字，有 30 个辅音字母和 4 个元音字母[1]，词汇由音节组成，音节之间由音节点 “·”（tsheg）隔开。因此，藏文分词是将连续的音节序列组合成词序列的过程，如藏语句字“ཚོང་སོག་/རྫུན་མ་/བཙོ་/བཏང་ཡེད་/།”（制销劣质产品，tsong sog rdzun ma bso vtsong byed）（本文用“/”表示藏语词语、音节间的分割符）。

藏文信息处理研究基础较为薄弱，分词研究大都是参考汉语的处理方法，结合藏文的实际情况，进行针对性的优化。藏文分词技术分类方法跟汉语分词技术分类方法基本相同，都可以分为基于规则方法和基于统计方法。基于规则方法需要词典支持，分词效果受词典影响很大，对未登录词和切分歧义处理能力较差，该研究时间长，研究成果也较为丰富。扎西次仁[2]所发表的“一个人机互助的藏文分词和词登录系统的设计”可以看作是藏语分词研究开始的标志。陈玉忠[3]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，该分词方案结合了藏文的特点，在一定程度上解决了切分歧义和未登录词问题，是一种较为有效的基

基金项目：国家自然科学基金（61032008, 61262054），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）。

作者简介：李亚超(1986－)，男，助教，主要研究领域为词法分析、机器翻译、机器学习、少数民族语言文字信息处理；加羊吉(1985－)，女，博士，副教授，主要研究领域为藏文信息处理；江静(1988－)，女，助教，主要研究领域为复杂网络。

于语言规则的分词方法。祁坤钰[4]提出了一体化的藏语三级切分体系，才智杰[5]提出了基于规则的方法“还原法”，来处理藏文分词中紧缩词识别问题，羊毛卓玛、欧珠[6]提出了一种改进型的藏文分词交集型歧义消解方法。以上研究是针对藏语语言特征，是借鉴汉语分词方法进行研究。

基于统计的藏文分词方法把分词问题看成序列标注问题，采用机器学习方法进行分类，最终得到分词结果。Liu [7]研究了基于分类模型的藏文数字识别，并且实现了基于序列标注的藏文分词方法[8]。史晓东[9]把基于隐马尔可夫模型的汉语分词系统 Segtag 移植到了藏文中，取得了 91%的准确率。江涛[10]实验了基于条件随机场的藏文分词方法，该方法把藏文按照音节进行切分，采用条件随机场的机器学习方法进行标注，取得了很好的效果。李亚超、宗成庆等[11]实现了基于条件随机场的藏文分词系统，并提出了自己的藏文音节标注系统，该方法处理了紧缩词问题，并把紧缩词识别和分词统一到一个模型中，在已知的音节标注系统中取得了最好的分词效果。Li [12]在四字位的标注集下，分别实现了基于条件随机场模型，最大熵模型，最大间隔 Markov 模型的藏文分词系统，并对实验结果进行对比。经过实验证明，藏文分词同样可以采用基于序列标注的方法，并且可以取得较好的分词效果。

前期，也有不少学者采用统计特征进行藏文分词研究，但只是采用如频率、熵等简单统计信息，这时期的统计特征只是作为基于规则分词方法的辅助。目前，基于序列标注的藏文分词方法采用的特征较少，大都采用音节位置，标点符号等特征，针对通过从无标注语料中抽取特征来提高有监督分词效果的研究较少。

藏文分词经过了十多年的研究，取得了较多的研究成果，但是目前仍然存在许多问题需要解决，并没有形成一个成熟的分词方法或者是共享的分词系统可以使用，分词仍然是制约藏文信息处理的瓶颈问题。

本文后续部分安排如下：第二部分详细介绍本文所采用的分词方法及分词特征选择，第三部分进行实验与分析，第四部分为本文的总结和下一步工作安排。

2 基于序列标注的藏文分词方法

2.1 藏文音节标注方法

基于序列标注的分词方法是汉语分词的主流方法，最新的藏文分词研究把该方法移植到藏文分词中，并取得了较好的效果。基于序列标注的藏文分词方法根据每个藏文音节在词中出现的位置，给予不同的标签，如图 1 所示。

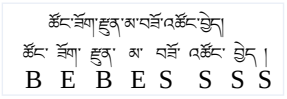

图 1 藏文分词标记示例

为了与文献 11 的分词效果进行对比，本文采用四音节位的标记集“BMES”。B、M、E、S 分别代表词的左边界、中间部分、右边界和单音节词，标记示例如表 1 所示，超过 3 音节的词中间部分都标记为 M。

表 1 音节标注示例

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་སྤྱུའུ་ཞི་(常务主席，rgyun las kruvu zhi)	རྒྱུན་/B ལས་/M སྤྱུའུ་/M ཞི་/E

紧缩词识别是为了确定藏文分词的基本单位，在现代藏语中，一些特殊的格助词与前面的音节之间没有音节点进行区分，这些音节的识别成为藏语分词中的一个重要研究内容。如“འདས་པ་ལོ་ལྔ་”（过去的五年，vdas pai lo lnga），第三个切分单位属格助词“ཨི་”和第二个切分单位“ལ་”之间没有音节点隔开。针对紧缩词识别，文献 5 和文献 11 分别提出了各自的识别方法。文献 5 提出的方法是一种基于规则的识别方法，需要词典支持，难以进行对比研究。文献 11 把紧缩词识别问题看成分类问题，这 6 个紧缩词按照功能进行划分，可以分为两大类，一类是作为格助词，另外一类是非格助词（包括基字和后加字），判断的依据为这些紧缩词的上下文特征[11]。这样，可以把紧缩词识别转化为序列标注问题，并且可以很方便的与基于序列标注的藏文分词结合在一起，并且在实验中取得了 98%以上的准确度，该方法作为本文的紧缩词识别方法。

本文提出的藏文分词系统最重要的特点是采用从无监督语料中抽取的特征，并将之融合到有监督的分词系统中，来增强传统的分词系统效果。本实验采用的特征分为两种，分别为基线特征和无监督特征，针对特征的选取在 2.2 和 2.3 部分进行详细介绍。

本文分词流程如下：首先，对输入的藏文文本进行预处理，紧缩词识别，得到分词基本单位；然后，采用条件随机场的序列标注的方法进行标注，由 CRF++[1]实现，根据标注的结果还原出初步的分词结果；最后，进行对初步的分词结果进行后续处理，得到最终的分词结果。分词系统流程图如图 2 所示。

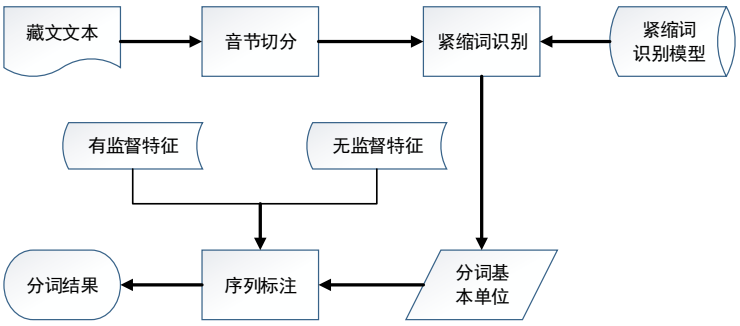

图 2 分词系统流程图

2.2 基线特征

在本文的基线系统中，采用文献 11 所采用的特征，包括音节位置特征和类别特征，音节位置特征包含了上下文特征，如表 2 所示。音节类别特征分为藏语音节、藏语标点符号、汉语标点符号、英文字母、英文数字、英文符号等。基线系统是进行无监督特征分词效果的对比系统。

[1] http://crfpp.googlecode.com/svn/trunk/doc/index.html

表 2 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

2.3 无监督特征

藏文分词缺乏大规模标注语料，从无标注语料中抽取特征，以提高有监督分词系统的效果是本文的研究目的。从无监督分词方法中受到启发，本文从无标注语料中抽取无监督特征，并将这些特征融合到基于序列标注的有监督藏文分词系统中，以此来提高本文基线分词系统效果，采用的无监督特征有以下几种。

2.3.1 边界熵

边界熵（Boundary Entropy, BE）作为无监督分词中判断切分边界的一个重要标准，广泛应用在汉语、英语等[13,14]语言的词语边界切分任务中。给定字符串 S=$C_{i..j}$,

$$h(x_{i..j}) = -\sum_{x\epsilon V} p(x|x_{i..j}) log p(x|x_{i..j}) \quad (1)$$

式(1)表示字符串 S 的边界熵，$h_L(C_{i..j})$和 $h_R(C_{i..j})$分别表示字符串 S 的左、右边界熵。熵是对事物不确定性的度量，熵越大不确定性就越大。如果一个字符串边界的熵变大了，那么该位置是词边界的可能性也较大。如果字符串 S 的左、右边界熵越大，那么该字符串有可能是一个完整的词。用 C_0 表示当前藏文音节，C_n 表示相对于当前音节的音节，本文抽取的字符串的边界熵如表 3 所示。

表 3 边界熵特征

特征	说明
$h(C_{0:-2})$	字符串 $C_{0:-2}$ 的边界熵
$h(C_{0:-1})$	字符串 $C_{0:-1}$ 的边界熵
$h(C_{0:1})$	字符串 $C_{0:1}$ 的边界熵
$h(C_{0:2})$	字符串 $C_{0:2}$ 的边界熵

2.3.2 邻接变化数

邻接变化数（Accessor Variety, AV）[15]表示一个字符串在上下文中的灵活程度，是对其在上下文中变化程度的度量。邻接变化数较大的字符串边界，该边界为分词切分边界的可能性也较大。即，一个字符串出现在不同的上下文环境中，那么该字符串成为词的概率也较大。

句 1：སྐུ་འབྲུམ་བྱམས་པ་སྦྲིང་གི་ཞེ་རྩྭ་མཆོད་པ། （塔尔寺的酥油花）；

句 2：ཞེ་རྩྭ་ལྷམ་རྡེའི་ནང་གི་གླན་ཀ （花园里的纠纷）；

句 3：ཁད་ཀྱིས་བཅོམ་པའི་ཞེ་རྩྭ （被霜凋零的花朵）；

句 4：གངས་ཀྱ་ཞེ་རྩྭ་ལས་ཚན། （雪莲花节目）。

对于“ཞེ་རྩྭ”（花朵）这个字符串，有四个前缀，分别为“གི་”，“B”，“འི་”，“ཀྱ་”，四个后缀，分别为“མཆོད་”，“ལྷམ་”，“E”，“ལས་”，其中“B”和“E”分别表示句子开始和结尾。AV 值为前缀和后缀的最小值，即一个字符串的 AV 值为 $AV(S)=\min\{L_{av}(S), R_{av}(S)\}$，$L_{av}(S)$表示一个字符串的左邻接变化数，$R_{av}(S)$表示一个字符串的右邻接变化数。

本文从藏文文本中抽取长度为 2、3、4 的字符串的 AV 值，作为无监督统计特征来增强有监督分词系统的效果。抽取特征如表 4 所示。

表 4 邻接变化数特征

特征	说明
$AV(C_{0:-3})$	字符串 $C_{0:-3}$ 的 AV 值
$AV(C_{0:-2})$	字符串 $C_{0:-2}$ 的 AV 值
$AV(C_{0:-1})$	字符串 $C_{0:-1}$ 的 AV 值
$AV(C_{0:1})$	字符串 $C_{0:1}$ 的 AV 值
$AV(C_{0:2})$	字符串 $C_{0:2}$ 的 AV 值
$AV(C_{0:3})$	字符串 $C_{0:3}$ 的 AV 值

2.3.3 无监督间隔标注

Voting Experts(VE)算法是一种局部最优的贪心算法，算法基于以下理论：相对的词内部熵较低，词边界熵较高，决定词的边界只需要局部信息[16]。VE 算法采用“专家”投票方式决定是否支持切分当前的序列边界。如果支持，当前边界的切分可能性增加。本文算法有两个“专家”，一个对于序列内部熵(Internal Entropy)较低的边界支持切分，计算公式为 $H_1(seq)=-logP(seq)$，seq 表示切分序列，另外一个对于序列边界熵较高的边界予以支持切分。

对于一个长度为 k 的待分词序列，在分词中有 $k-1$ 个位置需要决定是否切分。因此，整个切分序列的切分可能性有 2^{k-1} 个，这样算法复杂度较高，很难应用在实际分词应用中。VE 算法采用一种贪心算法，通过 $k-1$ 次计算就可以切分整个序列。

VE 算法在长度为 k 的待分词序列上，采用一个宽度为 $n(n<k)$的窗口，按照一定的顺序滑动到每个切分位置，对窗口覆盖的每个位置进行投票。如果有 m 个“专家”，那么每个位置会得到 $m*n$ 次投票。从上述分析可知 VE 算法时间复杂度为 $O(m*n*(k-1))$。我们在实验时发现，藏文分词中 n 为 3 时效果最好，本实验的 VE 算法中有两个“专家”，因此 m 为 2。通常 $m*n$ 是较小的整数。在实际分词中 k 通常远远大于整数 $m*n$，因此可以近似认为 VE 算法以线性时间运行。

用 $S=C_{1:k}=C_1C_2\ldots C_k$ 表示一个需要切分的藏文音节序列，用 $G_i(i=1\ldots k-1)$表示序列中 C_i 和 C_{i+1} 之间的间隔。每一个间隔 G_i 决定了 C_i 和 C_{i+1} 之间是否需要切分，这样可以把分词问题转化为序列间切分问题，通过采用相关的算法计算出间隔序列 $G_{i:k-1}$。本文用无监督的 VE 算法从无标注语料中计算每个间隔 C_i 的值，对于超过设定阈值间隔 G_i 予以切分，并将其融合到有监督的模型中，作为无监督特征的一种。

3 实验设置与分析

本实验中，切分的藏文分词语料题材为藏语小学语文课本，本语料由西北民族大学中国民族信息技术研究院组织人工标注。把整体语料分为测试语料和训练语料，训练语料包含 93563 个词，测试语料包含 17767 个词，测试语料未登录词比例为 5.6%。抽取无监督特征的语料题材为藏语文初、高中课本语料，包含 72 万个音节。

在本文的实验中采用的实验特征如下：实验 1，采用基线特征实现的藏文分词系统；实验 2，采用基线特征、边界熵实现的系统；实验 3，采用基线特征、邻接变化数实现的系统；实验 4，采用基线特征、边界熵、邻接变化数和无监督间隔标注实现的系统。

本实验把藏文音节串的边界熵、邻接变化数按照数值分为不同的类别，如表 5 所示，对于边界熵值取整数。

表 5 边界熵、邻接变化数分类表

AV 值类别	取值范围	边界熵类别	取值范围
A	1-3	A1	0-2
B	4-6	B1	3-5
C	7-9	C1	6-8
D	10-12	D1	9-11
E	13-15	E1	>11
F	>15	-	-

对于间隔标注阈值设定，本文经过实验选取 4， 即对于间隔标注值大于 4 的音节边界予以标记为切分边界。

下文 R、P、F、R_{OOV}、R_{IV} 分别表示召回率、正确率、F 值、未登录词召回率和登录词召回率，以此作为评测分析系统效果的指标，R、P、F 计算方法如下：

$$R = \frac{\text{正确切分的词数}}{\text{标准文本总词数}} \times 100\% \tag{2}$$

$$P = \frac{\text{正确切分的词数}}{\text{切分的总词数}} \times 100\% \tag{3}$$

$$F = \frac{2 \times R \times P}{R + P} \times 100\% \tag{4}$$

表 6 实验结果

实验	P/%	R/%	F/%	R_{IV}/%	R_{OOV}/%
实验 1	92.39	93.31	92.85	95.23	60.96
实验 2	93.89	93.27	93.58	94.32	75.36
实验 3	93.79	93.17	93.48	94.32	73.66
实验 4	94.13	93.50	93.82	94.38	78.74

从表 6 可以看出，与采用基线特征的分词系统相比，融合无监督特征的分词系统的各项指标均得到较大的提高，说明文本提出的从无标注语料中抽取的特征较为有效，可以明显提高基线系统的分词效果。

从实验 4 可以看出来，融合无监督间隔标注的藏文分词系统的未登录词召回率有了较大的提高，说明无监督间隔标注特征对于未登录词识别有较好的效果，并且与基线系统的分词系统相比，分词效果有了较大的提高。

与文献 11 相比，本文的分词系统整体效果较差的原因是本文实验的语料整体上较少，相当于前者的 8.5%，另外，未登录词比例也高于前者，因此本文实验结果整体上偏低。

4 结论与下一步工作

本文研究了从无标注藏文语料中抽取边界熵、邻接变化数、无监督间隔标注等特征，并融入了有监督的序列标注藏文分词系统中。实验结果表明本文抽取的无监督特征可以显著提高基线藏文分词系统的效果，并且可以很好的和有监督分词模型结合在一起。在后续的研究中，本文将在有监督的藏文分词系统中融合更加丰富的无监督特征，提高传统藏文分词系统的效果及分词系统的领域适应性，并研究无监督的藏文分词方法，以及资源受限条件下的藏文分词方法。

参考文献

[1] 山木旦，郑绍功，扎喜拉旦等．新编藏文字典[M]．西宁：青海民族出版社，1979．

[2] 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．

[3] 陈玉忠，李保利，俞士汶等．基于格助词和连续特征的藏文自动分词方案[J]．语言文字应用，2003，(1)：75-82．

[4] 祁坤钰．信息处理用藏文自动分词研究[J]．西北民族大学学报（哲学社会科学版），2006，(4)：92-97．

[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] 羊毛卓玛，欧珠[J]．一种改进型的藏文分词交集型歧义消解方法．西藏科技信息，2012,1:66-68.

[7] Huidan Liu, Weina Zhao, Minghua Nuo, Li Jiang, Jian Wu, Yeping He. Tibetan Number Identification Based on Classification of Number Components in Tibetan Word Segmentation[C]. In Proceedings of the 23rd International Conference on Computational Linguistics (Posters Volume) (Coling 2010), 2010:719-724.

[8] Huidan Liu, Minghua Nuo, Longlong Ma, Jian Wu and Yeping He. Tibetan Word Segmentation as Syllable Tagging Using Conditional Random Fields[C]. In Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation (PACLIC-2011), 2011:168-177.

[9] 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.

[10] Tao Jiang, Hongzhi Yu, Yangkyi Jam. Tibetan word segmentation system based on conditional random fields[C]. Software Engineering and Service Science (ICSESS), 2011 IEEE 2nd International Conference: 15-17 July 2011, 446-448.

[11] 李亚超，加羊吉，宗成庆，于洪志. 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，2013,4(27):52-58.

[12] Yachao Li，Hongzhi Yu. Study on Tibetan Word Segmentation as Syllable Tagging[C]. Natural Language Processing and Chinese Computing (NLP&CC 2013). 2013, 11: 363-369.

[13] Paul Cohen, Brent Heeringa, and Niall Adams. An unsupervised algorithm for segmenting categorical timeseries into episodes [J]. Pattern Detection and Discovery. 2002:117–133.

[14] Kumiko Tanaka-Ishii and Zhihui Jin. From phoneme to morpheme: Another verification rsing a corpus[C]. In Proceedings of the 21st International Conference on Computer Processing of Oriental Languages. 2011: 234–244.

[15] Haodi Feng, Kang Chen, Xiaotie Deng, and Weiming Zheng. Accessor variety criteria for Chinese word extraction. Computational Linguistics [J]. 2004, 30(1): 75–93.

[16] Paul Cohen, Brent Heeringa, and Niall Adams. An unsupervised algorithm for segmenting categorical timeseries into episodes[C]. Pattern Detection and Discovery. 2002:117–133.