文章编号：1003-0077(2020)02-0033-05

基于词性约束的藏文分词策略与算法

才让卓玛[1,2,3]，才智杰[1,2,3]

（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：自动分词作为自然语言处理基础性的研究课题，一直被学术界所关注，随着藏语自然语言处理技术研究的不断深入，藏文分词也面临越来越多的挑战。该文通过分析藏文自动分词研究现状，提出基于词性约束的藏文分词策略与算法。相对于传统方法，该方法不仅能有效地预防和处理各类歧义现象，而且在藏文未登录词处理方面有较好表现。

关键词：分词；词性；未登录词；歧义

中图分类号：TP391 **文献标识码**：A

Tibetan Word Segmentation Based on POS

CAI Rangzhuoma[1,2,3], CAI Zhijie[1,2,3]

(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, Xining, Qinghai 810008, China;
3. Key Laboratory of Tibetan Information Processing, Ministry of Education, Xining, Qinghai 810008, China)

Abstract: The word segmentation is a classical topic of natural language processing. This paper proposes a strategy and algorithm for Tibetan word segmentation based on the rules of part-of-speech. Compared with traditional methods, this method can not only effectively deal with the ambiguity, but also achieves better performance in processing unknown words in Tibetan.

Keywords: word segmentation; part-of-speech(POS); unknown word; ambiguity

0 引言

自动分词是信息提取、信息检索、机器翻译、文本分类、自动文摘、语音识别、自然语言理解等中文信息处理领域的基础研究课题[1]。藏文与中文一样，词与词之间没有任何显式分隔标记，因此自动分词是藏文信息处理的基础性问题。自 1999 年以来，藏文分词有了长足的发展，各种分词策略与算法涌现。1999 年，罗秉芬、江荻等[2]报道了藏语分词原则，扎西次仁[3]设计了一个藏文分词与词登录系统。2003 年，陈玉忠等[4]基于格助词和接续特征提出书面藏文自动分词方案，江荻[5]提出现代藏语组块分词方法。2006 年，祁坤钰[6]给出面向信息处理用一体化藏语分词方法。2009 年，才智杰等[7-9]基于最大匹配方法与格助词分块相融合的分词方案，提出用“还原法”和“碎片合并”方式解决藏文紧缩和未登录词。孙媛等[10]提出用双向最大匹配法和词频信息进行消歧的藏文分词方法。2011 年，史晓东等[11]通过移植汉语分词系统 Segtag 提出基于 HMM 的藏文分词方法。2012 年，扎西加等[12]提出基于条件随机场的藏文文本分词赋码一体化方案；

收稿日期：2019-08-15 **定稿日期**：2019-10-07

基金项目：国家自然科学基金(61966031, 61866032, 61262051)；国家社会科学基金(16BYY167)；教育部“春晖计划”(Z2016077, Z2012093)；青海省科技项目(2019-SF-129, 2017-ZJ-767)；青海省重点实验室项目(2013-Z-Y17, 2014-Z-Y32, 2015-Z-Y03)；教育部藏文信息处理与机器翻译重点实验室(2013-Y-17)

刘汇丹等[13]使用最大匹配方法与格助词分块进行分词，并在 1 000 句的测试集上进行了系统评测。2013 年，李亚超等[14]提出基于条件随机场的藏文分词方法。2014 年，康才畯[15]基于条件随机场模型，研究了统计模型与规则相结合的藏文分词方法。2015 年，李亚超等[16]基于条件随机场模型研究了基于音节标注的藏文分词。2018 年，李博涵等[17]以法律文本、政府公文、新闻为分词语料，研究了基于深度学习的藏文分词方法。

随着中文信息处理技术的迅速发展，藏文分词技术也取得了长足进步，但紧缩词识别、歧义处理和未登录词依然是藏文分词亟待解决的技术难题[18]。鉴于文献[19]的分词方法在 85 万字节语料上的分词准确率达 99%，标注准确率达 97%，本文参照文献[19]的词典结构、分词与标注规范，以紧缩词识别、消歧字段处理和未登录词识别为目标，研究基于词性约束的藏文分词算法。

1 分词策略与算法

1.1 问题描述

分词作为藏文信息处理基础性问题，存在紧缩词识别、消歧字段处理和未登录词识别等难题。藏文紧缩词是由于部分虚词（例如，属格འི、具格ས、为格ར、饰集词འང、离合词འམ、终结词ངོ）黏附于其前一个音节，使得两个词在书写形式上紧缩为一个音节，因而被称作紧缩词[7-8]。紧缩词的产生模糊了音节之间的分隔标志（音节符“·”），使得与非黏写形式的词同形。例如，句子“ཚང་མར་ཤོད་དང་།（大家说说）”中，为格“ར”黏附在音节“མ”后，形成与名词“མར”（酥油）、方位词“མར”（下）同形的黏写形式。

藏文文本中包含歧义词，包括交集型歧义和组合型歧义。例如，句子“ཐང་ཆེན་དུ་མ་མོ་མང་།（大地上有很多羊）”中，字“མ”和前一个字“ད”结合可组成词“ད་མ”，与后一个字“མོ”可结合成词“མ་མོ”，是一种交集型歧义；句子“རྒྱ་ནག་ན་ཚ་ཆེ།（南方很热）”中，“ན”可作虚词，也可与“ཚ”组成词“ན་ཚ”，是一种组合型歧义。歧义字段处理直接影响着分词系统的切分精度。

藏文文本中包含大量的未登录词，例如，人名（ལ་ཡུས，泰国人名）、地名（ཀླུང་ནན་ཉེ་ཇོ་ཀོང་ཀེ，中南海紫光阁）、组织机构名（ཅི་ཨར་ཅི་ནི་སྦྱི་མཐུན་རྒྱལ་ཁབ，吉尔吉斯共和国）等中外专有名词，又如不断涌现的网络词汇（མངོངས་མ，超女）或术语（བོད་ཁྱག，博客）。由此可见，未登录词的处理对分词效率的影响举足轻重。

1.2 分词策略

针对藏文黏写形式，本文将导致黏写形式的虚词集分成{ངོ}与{འི，འམ，འང，ས，ར}两类，根据词性约束分别处理。对于最大匹配方法无法检测出的歧义字段，通过考察每种词类与其他词类的互斥与接续的词性约束（例如，量词只能接续数词，具格助词接续动词，名词可接格助词、方位词等）建立规则库，进行预防和纠正歧义现象。对于未登录词，通过设置动态词典，将不在分词词典中的字符串存入动态词典，通过判断动态词典中字符串的频次是否大于给定阈值 ξ 分别进行处理。

特别地，考虑到兼类词是产生切分歧义的一个重要原因，且兼类情况复杂，如实词兼虚词、实词兼实词、虚词兼虚词。例如，不同语境下词“འབས”可作名词“命运”、动词“做”；“དུ”可作名词“烟雾”、疑问词“多少”、la 类格助词等。为了减少分词时的搜索代价，并提高消歧能力，本文利用不同词性在语料中出现的频率对兼类词进行了排序。

综上所述，本文分词算法基本框架设计如图 1 所示。

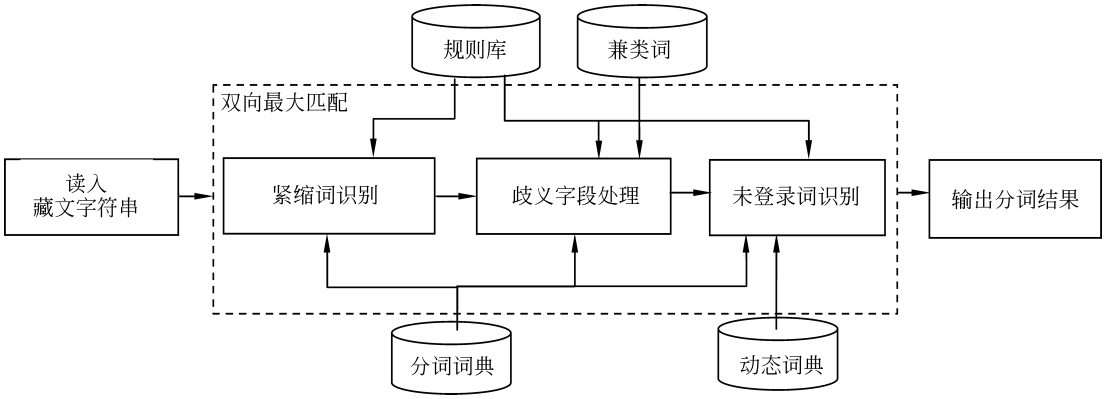

图 1 基于词性约束的藏文分词框架

1.3　分词算法

本文分词的基本思想是将词与词之间的互斥与接续规则作为词性约束条件提高算法的性能，并通过对兼类词进行排序、分类处理紧缩词、设置动态词典提高对紧缩词、未登录词与歧义字段的处理。算法中分词词典用 D 表示，动态词典用 D^* 表示，$\gamma\in$ {ང, ད, འམ, འང, ས, ར}。下面给出紧缩词识别、消歧和未登录词识别的核心算法。

算法 1　紧缩词识别

输入：串 $W_i=c_{i1}c_{i2}\cdots c_{in}$，其中，$c_{ij}$ 是串 W_i 的第 j 个字。

输出：串 W_i 的分词结果 seg。

步骤 1：判断串 W_i 是否包含 γ。若 $\gamma\in W_i$ 则将 γ 从 W_i 切出，使得 $W_i=c_{i1}c_{i2}\cdots c_{ij-1}+\gamma+c_{ij+1}\cdots c_{in}$，执行步骤 2（即 $\gamma=c_{ij}$）；否则读入下一串 W_{i+1}。

步骤 2：判断 $c_{i1}c_{i2}\cdots c_{ij-1}$ 及 $c_{i1}c_{i2}\cdots c_{ij-1}+$ འ是否在分词词典 D 中。若 $c_{i1}c_{i2}\cdots c_{ij-1}\in D$，则令 Temp=$c_{i1}c_{i2}\cdots c_{ij-1}$/+$\gamma$/，执行步骤 4；若 $c_{i1}c_{i2}\cdots c_{ij-1}+$ འ$\in D$，则令 Temp= $c_{i1}c_{i2}\cdots c_{ij-1}+$ འ/+γ/，执行步骤 4；否则执行步骤 3。

步骤 3：识别未登录词 $c_{i1}c_{i2}\cdots c_{ij-1}$，并且 seg=$c_{i1}c_{i2}\cdots c_{ij-1}$ +/+γ+seg_end，(seg_end 表示串 $c_{ij+1}\cdots c_{in}$ 的分词结果)，结束算法。

步骤 4：判断 c_{i-2}、c_{i-1}、c_{i+1}、c_{i+2} 是否满足 γ 的语法约束条件。若满足，则 seg=Temp+seg_end；否则串 W_i 中的 γ 为非紧缩词，seg 为串 $W_i=c_{i1}c_{i2}\cdots c_{in}$ 无紧缩词时的分词结果。

算法 1 首先判断串 W_i 是否包含“ང”或“ད”或“འམ”或“འང”或“ས”或“ར”，如果不包含，则说明该串中无紧缩词；否则串 W_i 可能存在紧缩词，此时利用分词词典 D 和紧缩词约束规则识别紧缩词，并获得相应的分词结果。例如，ཆེད་切分为ཆེ/ད་，དགའོ་切分为དགའ་/ཨོ་。

算法 2　歧义字段处理

输入：歧义串 W_i。

输出：串 W_i 的分词结果 seg。

步骤 1：判断串 W_i 中是否包含兼类词，若包含兼类词则将 W_i 切分为子串 W_{i1} 与 W_{i2}，兼类词包含于子串 W_{i1} 或子串 W_{i2}；否则读入下一串 W_{i+1}。

步骤 2：判断子串 W_{i1} 与子串 W_{i2} 是否满足语法约束条件，满足则输出切分结果 seg=W_{i1}/W_{i2}；否则进行下一步。

步骤 3：判断 W_{i2} 与 W_{i+1} 是否满足语法约束条件，满足则输出分词结果 seg=W_{i2}/W_{i+1}；否则下一步。

步骤 4：对比正、逆向匹配结果，对具有不同切分结果的歧义字段（如字段 S^* 被切分成 $W_{11}W_{12}$ 与 $W_{21}W_{22}$），由切分结果在词典中频率高低确定最终的分词结果。若 $W_{11}W_{12}$ 出现的频率大于 $W_{21}W_{22}$ 出现的频率，则 seg= W_{11}/W_{12}，否则 seg= W_{21}/W_{22}。

例如，“ཐང་ཆེན་དུ་མ་མོ་མང་།（大地上有很多羊）”可能被切分为“ཐང་ཆེན་/དུ་མ་/མོ་/མང་/།”或“ཐང་ཆེན་/དུ་/མ་མོ་/མང་/།”，根据虚词“དུ”的频次及与名词“ཐང་ཆེན”的语法接续特性将其正确切分为“ཐང་ཆེན་/དུ་/མ་མོ་/མང་/།”。

算法 3　未登录词识别

输入：不在分词词典 D 中的串 $W_i=c_{i1}c_{i2}\cdots c_{in}$，其中，$c_{ij}$ 是串 W_i 的第 j 个字。

输出：串 W_i 的分词结果 seg。

步骤 1：将 $W_i=c_{i1}c_{i2}\cdots c_{in}$ 逐字切分为 $W_{i*}=c_{i1}/c_{i2}/\cdots c_{in}/$。

步骤 2：判断串 W_{i*} 是否在动态词典 D^* 中，若不存在则将 W_{i*} 存入动态词典 D^*，并将其频次置为 1，seg=$c_{i1}/c_{i2}/\cdots c_{in}/$。

步骤 3：判断动态词典 D^* 中串 W_{i*} 的频次是否大于阈值 ξ，若是则将该串 W_i 作为一个专用名词进行切分，seg=$c_{i1}c_{i2}\cdots c_{in}/$；否则 seg=$c_{i1}/c_{i2}/\cdots c_{in}/$，并将其频次增加 1。

步骤 4：判断 D^* 中串之间是否满足接续条件，对于具有两种不同切分结果的字段，由这两种切分结果在词典中的频率高低确定最终的分词结果。

整篇文档分词完成后，对逐字分词的结果可利用动态词典进行再次处理，通过整体文档信息后处理没能识别的未登录词。

例如，将不在分词词典 D 的串“ཀླུང་ནན་ཏེ་ཇོ་ཀོང་ཀེ（中南海紫光阁）”切分成为“ཀླུང་/ནན་/ཏེ་/ཇོ་/ཀོང་/ཀེ/”后放入动态词典 D^*，并进行记数。若篇章分词后，串“ཀླུང་/ནན་/ཏེ་/ཇོ་/ཀོང་/ཀེ/”的频次大于 ξ，则将该串识别为一个专用名词。

2　实验及数据分析

本文所使用的分词词典共收录了 95 970 条词，

词条源于中小学藏文教材、期刊杂志及新闻语料，内容涵盖政治、经济、教育、文化及宗教等方面。实验语料共12 400句，其中8 400句筛选于小学藏语文教材(人教版)，3 000句摘自青海藏语网络广播电视台网站(http://www.qhtb.cn)新闻，其余1 000句摘自期刊和杂志。

本文将12 400句语料分成了训练语料与测试语料两部分，其中训练语料10 000句，测试语料2 400句。用分词算法对训练语料进行自动切分后，针对存在的问题进一步完善了分词词典与分词规则。将测试语料2 400句用本文算法切分后同人工分词结果进行了对比，测试语料中的紧缩词、未登录词、歧义字段统计数据见表1。

由于采用词性约束和兼容词的频次进行分词，本文算法对语料中出现的共307个紧缩词构成的黏写形式切分正确，歧义现象也得到了较好的解决，但对口语化文本中仅出现1次的未登录词依然出现识别错误现象。例如，“ གོན་ཀུང་ཞུ་ཏུའུ་ནས་ཚས་ཏེ་རྒ་རྒྱུ་རབ་ཁར་མ་སྒེབས་བར་ལ་མི་རྒྱང་ཀ་རྒྱང་དུ་ཁྱོན་བསྒོམས་འགག་སྐོ་ཐ་གཏོར་ཞིང་དམག་དཔོན་བཅུ་དྲུག་བསད་པ་རེད།(关公从许都出发到黄河渡口，千里走单骑，所历关隘五处，斩将六员)”(摘自《三国演义》译本《རྒྱལ་ཁབ་གསུམ་གྱི་གཏམ་རྒྱུད》)，分词时由于“གོན་ཀུང(关公)”和“ཞུ་ཏུའུ(许都)”两个词间缺失具格助词“གིས”而将其切成了一个词(གོན་ཀུང་ཞུ་ཏུའུ་/ནས་/)。

表1 测试语料中的紧缩词、未登录词、歧义字段统计表

类型	个数	示例	备注
紧缩词	68	ནམ་མཁའི་སྐར་མ།	འི(属格，འཇལ་ཐུ)
	18	ཡིན་ནའང་།	འང(饰格，རྒྱ་ཐུད)
	6	མིག་གམ་རྩ་བའམ་ཆ་ཤེའམ་ལུས།	འམ(离合词，འཇིད་ཐུད)
	15	དབང་བོ་ཐུབོ།	བོ(终结词，འཛོགས་ཚོག)
	76	ངས་གསར་འཕྲུར་དེ་གོ་ཐུང་།	_ས(具格助词，ཅིད་ཐུ)
	124	སྐད་ཆ་མདུན་དུ་མི་བཞད་པར།	_ར(la类格助词，ལ་དོན)
未登录词	15	སྨ་མར་པ་སི(Malpass)	
歧义字段	13	རྒྱ་ནག་ན་ཚ་ཆེའོ།	～表示潜在歧义字段

3 结束语

本文基于词性约束，通过兼类词排序、分类处理黏写形式对传统分词算法进行了优化。实验数据表明，该方法很大程度上避免了组合型歧义与交集型歧义的产生，提高了紧缩词识别效果。同时，通过设置动态词典，提高了未登录词的识别正确率。随着BBS、QQ(群)、BLOG(博客)、微信等各种网络交流平台的不断涌现，表达思想与情感的方式和工具越来越多样化，人们在各种平台上交流的语言也相应变得随意、简约和多样化，这无疑对藏文自动分词又提出了新的挑战。

参考文献

[1] 闻玉彪，贾时银，邓世昆，等.一种改进的最大匹配中文分词算法[J].计算机技术与发展，2011，11(10)：92-95.

[2] 罗秉芬，江荻.藏文计算机自动分词的基本规则//民族语文现代化论集[M].北京：民族出版社，1999.

[3] 扎西次仁.一个人机互助的藏文分词和词登录系统设计.中国少数民族语言文字现代化文集[M].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，2003：75-82.

[5] 江荻.现代藏语组块分词方法与过程[J].民族语言，2003(4)：31-39.

[6] 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，2006(4)：92-97.

[7] 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，2010，21(2)：75-77.

[8] 才智杰.藏文自动分词中紧缩词识别[J].中文信息学报，2009，23(1)：35-38.

[9] 才智杰，才让卓玛.藏文自动分词系统的设计[J].计算机工程与科学，2011，33(5)：151-155.

[10] 孙媛，罗桑强巴，杨锐，等.藏语自动分词方案的设计[C].中国少数民族语言文字信息处理研究与进展.北京：民族出版社，2009：228-237.

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

[12] 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，2012，27(1)：57-61

[13] 刘汇丹，诺明花，赵维纳，等.一个实用的藏文分词系统[J].中文信息学报，2012，26(1)：97-103.

[14] 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.
[15] 康才畯.藏文分词与词性标注研究[D].上海：上海师范大学博士学位论文，2014.
[16] 李亚超，江静，加羊吉，等.一个开源的藏文分词词性标注系统[J].中文信息学报，2015，29(6)：203-207.
[17] 李博涵，刘汇丹，龙从军，等.基于深度学习的藏文分词方法[J].计算机工程与设计，2018，39(1)：194-198.
[18] 才让卓玛，基于混合基元的藏语语音合成技术研究[D].西安：陕西师范大学博士学位论文，2016.
[19] 才智杰，才让卓玛.班智达藏文标注词典设计[J].中文信息学报，2010，24(5)：46-50.

才让卓玛(1970－)，博士，教授，主要研究领域为人机语音交互、藏文信息处理。
E-mail：cr-zhuoma@163.com

才智杰(1970—)，博士，教授，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

(上接第 32 页)

[16] Mikolov T，Le Q V，Sutskever I.Exploiting similarities among languages for machine translation[J].arxiv.preprint arxiv：1309.4168，2013.
[17] 斯·劳格劳，内蒙古蒙科立蒙古文化股份有限公司.蒙古文自动校对(试用版)[CP/OL].[2018-11-05].http://mts.menksoft.com/home/Jindex.
[18] Sun M，Chen X，Zhang K，et al.Thulac：An efficient lexical analyzer for Chinese[R].Technical Report，2016.
[19] Kingma D P，Ba J.Adam：A method for stochastic optimization[J].arxiv preprint arxiv：1412.6980，2014.
[20] Srivastava N，Hinton G，Krizhevsky A，et al.Dropout：A simple way to prevent neural networks from overfitting[J].The Journal of Machine Learning Research，2014，15(1)：1929-1958.
[21] Hochreiter S，Schmidhuber J.Long short-term memory[J].Neural Computation，1997，9(8)：1735-1780.

曹宜超(1994—)，硕士研究生，主要研究领域为低资源神经机器翻译、数据挖掘。
E-mail：cycao@mail.ustc.edu.cn

李淼(1955—)，通信作者，研究员，博士生导师，主要研究领域为人工智能农业知识工程。
E-mail：mli@iim.ac.cn

高翊(1970—)，高级工程师，主要研究领域为少数民族文字—汉文智能翻译、农业知识工程、数据库等。
E-mail：498898209@qq.com