基于格助词和接续特征的藏文自动分词方案*

陈玉忠 李保利 俞士汶 兰措吉[1]

（北京大学计算语言研究所 北京 100871） （青海师范大学 西宁 810008）[1]

摘要： 本文结合藏文各类形态特征，首次提出了一种基于格助词和接续特征(BCCF, Based on Case-auxiliary word and Continuous Feature)的书面藏文自动分词方案。其总体技术特点是：在格助词、接续特征、字性知识库以及词典支持下，进行逐级定位的确定性分词。初步测试表明：这一方案在发现和消除切分歧义、解决未登录词问题，进而在提高藏文分词精度方面具有很高的实用价值。

关键词： 格助词；接续特征；藏文分词

1. 引言

随着对语言信息处理研究工作的不断深入，藏文信息处理技术也从文字处理逐步转向语言信息处理。与汉语、日语等语种的信息处理一样，藏文自动分词是藏文信息处理中一项不可缺少的基础性工作。书面藏文分词问题解决的好坏，直接制约着藏文词频统计工程、藏外机器翻译等高层藏文信息处理技术的进一步发展。因而，设计并实现实用化的书面藏文自动分词系统已势在必行。同时，结合藏文特点开展的分词研究，反过来对其他语言的分词研究也具有非常重要的参考价值。

藏文分词研究目前是一片空白，借鉴亲属语言汉语分词研究的已有成果和成功经验，无疑对把握分词问题的本质，针对性地开展藏文分词研究有重要的指导意义。迄今为止，汉语分词方面已提出了许多极有价值的分词方案[2]。从所采用的分词方法来看，这些方案大致上可分为两大类，即统计方法和规则方法。统计方法是先建立一个自动分词统计模型[3]，获取模型的各组参数，然后从各种可能的词串中挑选概率最高的词串作为输出结果。而规则方法是利用词表和规则，采用一定的算法[①]，拿文本中的候选词去跟词表中的词匹配。匹配成功且符合规则要求，则将候选词确定为词并予以切分输出。不论是统计方法还是规则方法都存在两大难题：一是歧义切分问题，二是未登录词问题[4]。通过对不同方案的对比分析使我们认识到，不同的分词方案模拟了人类分词行为的不同侧面，都存在各自的切分盲点，也就是说分词精度与分词方案有关。为此，我们认为，在目前的情况下，藏文分词的首要任务是研究并提出符合藏文特性的最佳分词方案，以提高分词系统的切分精度和通用性，进而尽可能地逼近人们所期望的各类分词应用需求。

衡量实用化书面藏语分词系统的关键指标是系统切分精度[②]。切分精度通常以切分正确率来衡量[1]，切分正确率愈高表示切分精度愈高，反之亦然。由于分词精度直接关乎系统的正确性、科学性以及整体

* 本文相关研究得到国家 863 计划（2001AA114040）和 973 项目（G1998030507-4）资助。

[1] 现有汉语分词系统主要采用最大匹配算法，即 MM 法(The Maximum Matching method)。有关分词算法的详细介绍参见文献[2]。

[②] 切分正确率的定义见本文第五部分。通常切分速度也是衡量分词系统性能指标的一项不可忽视的指标。但随着计算机软硬件技术的飞速发展，提高分词速度已变得越来越容易。因此，本文未将它纳入评价系统的性能指标中。

性能，因此，提高切分精度自然就成为了整个藏文分词系统设计、实现过程中的一个关键点和核心问题。

本文包括如下几部分：第二、三部分首先分析了藏文文本自动分词的难点以及藏文文本的特点，在分析比较了两种基本的分词方法——最大匹配法和格助词分词法的基础上，第四部分提出了基于格助词和接续特征的书面藏文分词方案，最后给出了实验结果以及进一步工作的设想。

2．藏文文本自动切分的难点探析

自从80年代初中文信息领域提出自动分词以来，虽然经过有关方面的众多专家、学者为之付出了不懈的努力，但还未研制出一个与人们的期望相一致的通用的实用系统，这从一个侧面说明了自动分词问题所固有的复杂性。

结合藏文的特点，我们首次提出了一种基于格助词和接续特征[①]（BCCF，Based on Case-auxiliary word and Continuous Feature）为主的分词方案，所采用的方法本质上属于“规则+特征”的方法。为了说明这一方案的有效性和实用性，首先有必要对规则分词和格助词分词在藏文分词中引起的错误切分类型作一番实际考察。

2.1 规则分词法及其切分难点分析

规则分词通常采用最大匹配算法，其最大优点是算法简单、容易实现。那么，这一分词方法在藏文中会引起哪些歧义错误，存在哪些切分难点呢？我们通过对5900个词的（500句的综合语料）实际切分发现，采用这一方法后引起的切分错误共有750次，占语料总词数的12.71%。主要错误类型及其所占比例如下：

Ⅰ.交集型歧义[②]错误442次，占整个切分错误的58.93%，典型实例见(1)。其中“/”表示正确切分序列，“+”表示错误切分处，“+/”表示该处未能正确切分（下同）。

Ⅱ.组合型歧义错误212次，占整个切分错误的28.27%，实例见(2)。

Ⅲ.紧缩格（见下文说明）识别错误52次，占整个切分错误的6.93%，实例见(3)。

Ⅳ.未登录词切分错误44次，占整个切分错误的5.87%，实例见(4)。

ག་+དོན་/སྒྲ་/འཕྲད་པ། ཚོག་/སྒྲག་+ཅ/ཏེ་/ཐ་སྙད། ཅ་ཁང་/ནང་/ཏུ་/ཁ་+ཅ་/ཚད། (1)

ཁབ་/ཞིང་+ཚོགས་/ཀྱང་/ཚད། རྒྱ་/ནག་/ནང་+ཚོ་/ཆེ། གྱག་རྟགས་ལ་+རེ། (2)

སྔོང་/རུ+/ས་/འཁྱར། དེ་ཆ+ར་/འབད། རོག་/ཀྱ+/ས་/ཐུབས། (3)

ཞེན་+ཞེ་+ཐན་/གཤང་/ལྷགས། ཐེ་+བཏུན་+གྱག་/གི་/གནད་དོན། ཁ་+ཀེན་+ཟེན་/ཞི་/ཞི་རིགས། (4)

在此基础上，我们对引起以上四类错误切分的根源作了进一步考察后发现：

[①] 格助词和接续特征的介绍见本文第三部分。

[②] 藏文中交集型歧义字段、组合型歧义字段以及未登录词的定义与汉语完全相同，具体定义参见文献[2]。

Ⅰ类切分错误主要由实词-实词、实词-虚词、虚词-实词和虚词-虚词四类词的交集型字段产生。其中，实-虚和虚-实这两类交集性字段产生的错误可以用“词典+规则”的方法给出正确切分；虚-虚交集型字段产生的概率非常小，可用穷尽法解决；而实-实交集型字段产生的错误用规则分词法无法从根本上予以解决。

Ⅱ类切分错误也主要由实-实、实-虚、虚-实和虚-虚四类词的组合型歧义字段产生。其中，实-虚、虚-实和虚-虚类组合型歧义字段产生的错误可采用与Ⅰ雷同的方法解决；实-实组合型歧义字段产生的错误用规则分词法仍然无法解决。

Ⅲ类切分错误是指ར་ས་ནི་ཏ 四个藏文格助词与前置词紧缩（无分字点）结合引起的。其中后两个格助词引起的紧缩错误可以用“词典+规则”的方法解决；而前两个格助词引起的紧缩格识别错误用规则分词法解决起来代价太大。

Ⅳ类错误主要因词典收词有限引起的。一般来说，解决办法有二：一是扩大词典收词，但这会增加部分交集型和组合型歧义错误；二是引入统计方法辅助解决。由于新词术语不断涌现，在规则分词的情况下，这两种办法都不能从根本上解决未登录词的切分错误。

2.2 格助词分词法及其切分难点分析

对于藏文而言，利用格助词和接续特征分词，理论上有两大好处。首先，由于这种方法与词典无关，因而避开了未登录词问题。其次，词的切分转化为格助词及其接续特征的识别问题。但是，由于这一方法完全依赖于格助词，因而堆块（因词间无格助词，多个连续词未能切开）错误、格助词识别（兼类格和紧缩格）错误和截断（因词内出现格助词，把一个词切成多个词）错误自然就成了首先要解决的关键问题。我们利用格助词通过对上述同一语料的实际切分发现，分词错误共出现 664 次，占语料总词数(5900 个词）的 11.25%。其中：

Ⅴ.堆块错误占绝大多数，共出现 520 次，占整个切分错误的 78.32%；典型实例见(5)。斜杠和加号所代表的意义同上。

Ⅵ.格助词识别错误次之，共出现 104 次，占整个切分错误的 15.69%；典型实例见(6)。

Ⅶ.截断错误最少，共出现 40 次，占整个切分错误的 6.09%；典型实例见(7)。

ཀྱ་ཚོ་ལ་/གཞར་བ/ཡི་ཁེ་/འཇི་བ/　ཤད་ཡིག་/དེ་ཁ་ཤོད་/གདུགས་ཆེ/　(5)

ཡང་/ཁྱེ/　རམ་འདེགས/　ཏུ་རམ/　ཐུབ་ཏུ/　ཕྱི་ལང་/　(6)

འཛམ་ཁ་ནི་/དཔྱང་ས/　གནམ་ཕྱི་/ཐི་བདུན/　དབང་དུ་/བཏང་ན/　(7)

以上三类错误切分中，Ⅴ类错误主要由“名词+VP|NP|MP|AP”型和“DP+形容词”型堆块错误引起的，这类错误用格助词方法很难解决；Ⅵ类错误主要由实词、虚词兼类格识别错误和紧缩格识别错误

引起的，这类错误用“格助词+规则”可部分解决。Ⅶ类错误主要是由专名和复合关联词内出现的格助词引起的。其中，复合关联词错误可用穷尽法解决，专名切分错误无法用格助词法解决。

由以上讨论可见，仅仅利用词典进行规则分词，无法从根本上提高分词精度。单纯用格助词信息进行分词，最大的贡献在于把本来存在于句子级的切分歧义压缩到了短语一级，但短语内的词切分问题仍然没有得到解决。将以上两种方法加起来使用仍不能有效排除Ⅰ、Ⅱ、Ⅵ、Ⅶ（ⅠⅡⅥ包含了ⅢⅤ）类的一些关键性歧义错误。从系统论的观点来看，多种方法在一个系统中有机结合、优势互补，可以使整体效果达到最佳。这说明要有效地提高藏文分词精度，除了格助词信息外，还需要充分利用藏文其他各类特征信息。为此，我们先来分析一下藏文有哪些有效的形式特征可供分词使用，再来提出符合藏文特性的最佳的分词方案。

3．藏文文本的特点

为了探寻提高分词精度和通用度的有效途径，我们要尽可能地把对分词有用的特征信息挖掘出来[5]。从分词角度来看，藏语言文字的特征包括字切分特征、词切分特征和句切分特征(这些特征都服从一定的语法接续规则，本文统称为接续特征)。为此，我们有必要首先对书面藏文的这些接续特征有一个比较清楚的认识，只有在充分把握藏文各级切分特征的基础上，才有可能提出一套符合藏文特征的分词方案。

字切分特征 从藏文的文字特征来看，可利用的切分特征主要有以下几点[6]。

一是音节特征，藏文是拼音文字，她由30个辅音字母、4个元音字母以及上、下加字（辅音字母的变体）组成。藏文字以音节为单位，每个音节最少可由一个辅音字母构成（元音和上、下加字不能独立成字），最多可由7个字母拼合而成，各音节间用音节点分隔。

二是拼写特征，藏文自左向右书写，组成音节时以基字为中心分为前置字、后置字和又后置字（合称为后加字），基字可横向和纵向双向拼写，而前置字和后加字只能横向拼写。

三是形态特征，藏文由确定的10个辅音字母作后加字，其形态特征都发生在这10个确定的后加字上。

四是标点符号特征，藏文有一套独立而完整的标点符号体系，主要在篇章、段落、句子和字之间起“分界符”的作用。

词切分特征 藏语词从总体上分为实词和虚词两大类。藏语实词可单独使用且具有具体的词汇意义，它包括名词、代词、动词、形容词、数词等。虚词有格助词、关联词等，不能单独使用，只能粘着在实词后面起语法作用或表示某种逻辑关系[7]。从藏文词语的形态特征来看，明显的切分特征主要有以下几点：

一是格助词接续特征，藏文格助词（case-auxiliary word）的个数不多[①]，但使用频率极高。大多数藏文格助词在添接时，要严格按前一词（或字）后加字的粘着性形态变化规则添接。比如传统藏文文法规定：属格助词的粘着性形态有五种变体，根据接续规则，前趋实词的后置字不同则所接的属格助词就应不同。如例句（8）中下划线部分所示。

[①] 本文中，格助词包含参考文献[7]所列举并确认的所有八格和虚词中的语法虚词；下文中的关联词则指除格助词之外在词与词、句子与句子以及段落之间起关联作用的其他语法虚词。

སློབ་ཁང་གི་　　སློབ་དེབ་ཀྱི་　　དཀོན་ཁག་གི་　　སློབ་མའི་　　ཁང་པ་ཡི་　　（8）

教室的　　教材的　　教师的　　学生的　　房子的

二是动词的屈折形态变化，现代藏语只有动词还保留着时、式、态等曲折形态变化[8]。这是动词有别于其他词类的重要特征。

三是名物化词缀特征，藏文动词、形容词在句子中修饰名词性成分或作非谓成分时，一般都要进行名物化转换，即要添接名物化后缀。

四是重叠结构，藏文的重叠结构主要发生在形容词当中，常见的重叠形式有 AA 式、ABB 式、ABCB 式等三种。

五是动名词的动词性词缀特征，顾名思义藏文动名词兼有动词和名词两种语法功能，是藏文特有的一类词。其特点是通过后接固定的几个动词性词缀实现词性转化。

句切分特征 在任何语言中，句子结构都是规则有序的，而其构成要素之间又是相互制约的，这是人类语言的共性特征。但在不同语言中，句子结构又是各不相同的，这是不同语言个性特征的具体表现[9]。藏语句子结构的主要特征概括起来有以下三点：

一是语序特征，藏语是 SOV 型语言，即谓语动词后置型语言。动词是句子的核心，决定着格助词的添接类别。如：

藏语：ཚོ་ངེ་ང་གིས་（s）བོད་ཡིག་(o)སློབ་གྱིན་འདུག(v)　S+O+V　(9)

汉语：才让(S)正在学习(V)藏语(O)　S+V+O

二是主要借助格助词来表达句子含义的作格特征，藏语句子的主要成分一般都要与格助词相关联，只有这样才能把句子各成分之间的语义关系表达清楚。具体地说，藏语主语多与使动格助词相连，宾语多带有受动格助词，定语与中心语之间多以属格助词相联接，状语与谓语之间多用使动格助词和自性格助词相联接。比如：

སྒྲལ་མས་ནན་ཏན་གྱིས་དཔྱིན་ཇི་ལེ་ཚོང་ཚག་འབྲི་དེབ་སྟེང་བྱི་སྟ། (10)

汉语：卓玛 认真地 将英语 单词 写在 本子上。

三是藏语短语的后修饰特征，一般情况下，藏语形容词、数词、代词等与名词结合构成短语以及动词与助动词结合构成短语时，其中心语在前，修饰语在后。如：

藏语：ཚིང་ཚག་གསར་པ།　ཉུག　སོ་བརྒྱད།　དེ་ཚ་དེ།　སློབ་སྦྱོང་ཅེད་དགོས། （11）

汉语：新单词　六人　八元　那本书　应该学习

综上所述，充分利用藏文的上述接续特征来指导我们识别词与词或短语与短语之间的切分点，帮助我们消除部分切分歧义无疑是非常有帮助的。

4．基于格助词和接续特征的书面藏文分词方案

4.1 设计思想

由本文第二部分的讨论可见，无论单纯地用规则分词还是单纯地用格助词分词，都不能有效地提高分词精度。由本文第三部分关于藏文各类接续特征的分析使我们认识到，藏文在字、词、句各级存在着许多天然的切分特征标记。据此，本文提出利用字切分特征和字性库先“认字”，再用标点符号和关联词“断句”，用格助词“分块”，再用词典“认词”，充分利用各类接续特征“分词”的多级分词方案。其基

本处理流程如图 1 所示。

4.2 处理策略

由图 1 的处理流程可见，正确识别和处理好各个阶段的切分歧义是保证有效提高本方案切分精度的核心问题。本文以下主要从四个方面来论述本方案在不同阶段所采取的切分歧义识别策略。相应的排歧策略及其实现算法将另文介绍。

认字和断句是保证有效利用接续特征的基础 认字相对简单，可用分字点、标点符号和字性库来完成；断句则可利用标点符号和关联词来实现。由于藏文格助词都是单字词，动词和形容词(不计名物化后缀)也大都是单字词，句子内动词居尾且决定格助词的接续类别。通过这一步一方面把字性信息附加到每个字上，另一方面划清了句子的界限，从而为开展后续工作奠定了基础。

基于字切分特征和句切分特征的分块策略 分块的关键任务是正确识别格助词（Ⅵ类错误）。藏文格助词可分为无歧义格助词和歧义格助词两类，无歧义格助词通过认字就可确定；歧义格助词进一步可分为紧缩格和兼类格两类，其中紧缩格可采用字切分特征和句子特征识别；兼类格又包括规则兼类格和不规则兼类格两类，规则兼类格可用字性特征、字切分特征和句子切分特征来识别。不规则兼类格在这一步不能完全确定，但要加不可识别标记。

基于词典、格助词、字和句接续特征的认词策略 认词过程其实就是识别每个块是否是一个“可能”的词的过程。通过分块切分出来的单字块，只要词典中存在，原则上可认定为一个词(可能有截断错误)。而多字块则需要识别Ⅰ、Ⅱ类错误(交集性和组合性歧义)，对于Ⅰ类歧义可采用局部正向和反向最大匹配算法（MM 和 RMM）来识别；而Ⅱ类中，除实-实型组合歧义外，采用格助词、字和句接续特征予以识别；实-实型组合歧义目前主要用个性规则识别，无法识别时默认长词优先。

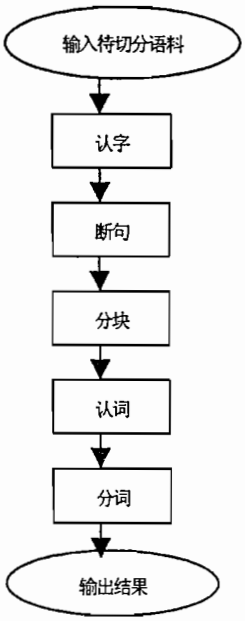

图 1 基本处理流程

基于字、词、句和格助词接续特征的分词策略 分词过程是整个流程的最后一步，也是本方案的关键性一步。这里要综合运用已有字词句的接续知识统一扫描一遍整句，来识别兼类格和Ⅶ类错误（截断错误）。其中，兼类格主要采用动词与格助词的一致性检查来识别；词截断歧义中复合接续词采用模式匹配加个性规则来识别。对于词典忠不存在的部分专名截断留待词性标注和句子分析阶段来处理。

4.3 切分精度

由本文第二部分讨论可知，在保证格助词正确识别的情况下，5900 个词中堆块错误共出现 520 次。假定交集和组合性歧义都出现在这 520 个当中，那么，我们就可以来合理地预测本方案的切分精度。按规则分词法在语料中引起的歧义比例（交集和组合性歧义分别占 7.5%和 3.6%）可知，这两种歧义在这 520 个堆块当中可能分别出现 39 次和 19 次。假定所有交集性和组合性歧义中实-实型歧义平均占四分之一，则可能的切分错误有 15 次；再假定词截断错误中专名截断占三分之一，即出现 15 次，以上两项合计 30 次，占总词数的 0.51%，约为

千分之五。即从理论上来看，本分词方案切分精度的下限为99.49%。基本上达到了实用化的要求。

4.4 系统的通用性

由上述处理策略可见，本方案的核心是基于格助词和接续特征的分词，词典只是辅助手段，即方案与词典的相关度很小。这就使得本分词方案对不同领域、不同内容（藏文格律体除外）的藏文语料将会表现出较强的适应性。

5. 实验与评价

迄今为止，我们尚未见到国内外有关藏文自动分词方案方面的文献报道。为了说明本分词方案对藏文分词的有效性，我们依据本方案设计实现了一个实验系统，并采用分词正确率（精度）作为指标来对方案作了初步的测试评价。

分词正确率=（切分结果正确词数/语料总词数）×100%　　（12）

语料来源：分别从新编藏文字典、藏文文法详解、青海教育（杂志）、青海日报（报纸）、高等数学中抽取100句。

评价方法：对同一语料分别用基于规则的方案、基于格助词的方案和基于BCCF的方案实现的子系统进行切分，对切分结果进行人工统计后得到如表1所示的结果。

表1 不同方案对比测试结果

分词方案	测试集(句)	测试集(词)	错误（词）	精度（%）
基于规则的方案	500	5900	750	87.29
基于格助词的方案	500	5900	664	88.74
基于BCCF的方案	500	5900	155	97.21

实验表明，上述三种方案中，基于BCCF的分词方案的切分正确率远远高于其它两种方案。同时，由于该方案摆脱了词典的束缚，不受领域限制因而具有较强的通用性。

6. 结束语

从长远研究考虑，通过对藏文字、词、句以及格助词等接续知识的深入研究，不仅有利于提高藏文分词精度，而且为进一步开展藏文词性标注、句法分析以及句子生成等重大研究课题开辟了一条极有价值的“通道”。为此，在此基础上，我们计划一方面扩大测试语料，细化接续特征信息，优化系统方案；另一方面引入统计技术，开展基于格助词和接续特征的藏文分词、词性标注和句子分析的一体化分析技术研究。

参 考 文 献

[1] 何克抗，徐晖，孙波. 书面汉语自动分词专家系统设计原理. 中文信息学报，1991，5(2)

[2] 刘源，谭强，沈旭昆. 信息处理用现代汉语分词规范及自动分词方法，北京：清华大学出版社，1994

[3] 孙茂松，黄昌宁等. 中文姓名的自动辨识. 中文信息学报，1995，9(2)

[4] 刘开瑛. 现代汉语自动分词评测技术研究. 语言文字应用，1997，21(1)

[5] 俞士汶，计算语言学的应用研究与基础研究. 辉煌二十年—中国中文信息学会二十周年学术会议论文集，北京：清华大学出版社，2001

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

[7] 才旦夏茸. 藏文文法详解，西宁：青海民族出版社，1988

[8] 嵌绕威色木，藏文动词释难，成都：四川民族出版社 1994

[9] 多识. 藏语语法深义明释，兰州：甘肃民族出版社，1999

[10] 刘开瑛，中文文本自动分词和标注，北京：商务印书馆，2000

作者简介： 陈玉忠（1963—），男，青海果洛人，博士生，副教授，主要研究领域为机器翻译、藏文信息处理；李保利(1971—)，男，河南洛阳人，博士生，主要研究领域为信息提取；俞士汶(1938—)，男，安徽人，教授，博士生导师，主要研究领域为计算语言学；兰措吉(1967—)，女，青海海南人，讲师，主要研究领域为现代汉语语法。

A Tibetan Segmentation Scheme Based on Case-auxiliary Word and Continuous Features

CHEN YuZhong[1] LI BaoLi[1] YU ShiWen[1] LAN Cuoji[2]

[1](Institute of Computational Linguistics, Peking University, Beijing 100871, China);
[2](QingHai Normal University, XiNing 810008, China)

E-mail: degai@pku.edu.cn

Abstract: A cascaded written Tibetan word segmentation scheme, which is based on Case-auxiliary word and Continuous Features, is proposed in this paper. Using inflectional morphology information such as case-auxiliary word and continuous feature, and adopting a cascaded strategy, are the key features of the proposed scheme. Preliminary experiments suggest that it could detect and eliminate segmentation ambiguities and deal with unknown words. The scheme has much more practical value, as is indicated in the higher precision.

Key words: Case-auxiliary Word; Continuous Feature; Tibetan Word Segmentation