藏文自动分词系统的设计与实现*

陈玉忠 李保利 俞士汶

{degai,libl,yusw}@pku.edu.cn http://icl.pku.edu.cn/

（北京大学计算语言学研究所 北京 100871）

摘要：藏文自动分词系统的研制目前在国内仍是空白。本文从四个方面详细报告了书面藏文自动分词系统的具体实现过程，内容包括系统结构、分词知识库的组织与实现以及分词策略、算法设计及其详细的自动分词过程实例。文章最后给出了实验结果，结果表明系统具有较高的切分精度和较好的通用性。

关键词：格助词、接续特征、藏文、自动分词

中图分类号：TP391

the Design and Implementation of a Tibetan Word Segmentation System

CHEN Yu-zhong LI Bao-li YU Shi-wen

(Institute of Computational Linguistics, Peking University, Beijing 100871, China);

Abstract: Word segmentation for Tibetan has not been well studied yet. This paper reports a Tibetan word segmentation system that we designed and implemented. Several issues about the system are explained, which include system architecture, knowledge bases, segmentation strategy, and algorithms. In preliminary experiments, the system demonstrates higher accuracy and domain independency.

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

1、引言

随着对语言文字信息处理研究工作的不断深入，藏文信息处理技术也从字信息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（Tibetan Automatic Word Segmentation）是藏文信息处理中一项不可缺少的基础性工作。

一般从处理过程来看，我们可以把自动分词看作是用计算机自动识别文本字符流中的词并在词与词之间加入明显切分标记[1]的过程。从应用需求来看，自动分词的主要目的是确定自然语言处理的基本分析单位，为进一步开展自动分析进而为实现机器翻译、篇章理解、自

* 本文研究工作得到国家自然科学基金项目（合同号：69663001）和 973 项目（合同号：G1998030507-4）资助，特此致谢。

作者 陈玉忠，男，1963 年生，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，男，1971 年生，博士生，主要研究领域为中文信息处理；俞士汶，男，1938 年生，教授，博士生导师，主要研究领域为计算语言学。

[1]英语、法语、俄语、德语等语言的词与词之间一般都采用自然的空格作为切分标记。所以，汉语、藏语和日语等语言的实际切分系统中，词与词之间通常也采用空格作为切分标记。本文为了明显起见，下文所有示例中藏文词与词之间的切分标记都采用左斜杠‘/’。

动文摘、文本校对、自动标引等应用处理系统做好前期准备工作。

本文依据基于格助词[1]和接续特征（Based on Case-auxiliary word and Continuous Feature，BCCF）的书面藏文自动分词方案中提出的总体设计思想[1]，设计并实现了一个基于格助词和接续特征[2]的书面藏文自动分词系统。主要分四个部分详细报告了书面藏文自动分词系统的具体实现过程。第二部分是系统总体结构，第三部分是系统各类知识库的组织与实现，第四部分是分词策略和算法以及详细的分词过程举例，最后是实验结果和结论。

2、系统结构

2．1 系统结构

基于格助词和接续特征的书面藏文自动分词系统的结构如图 1 所示。

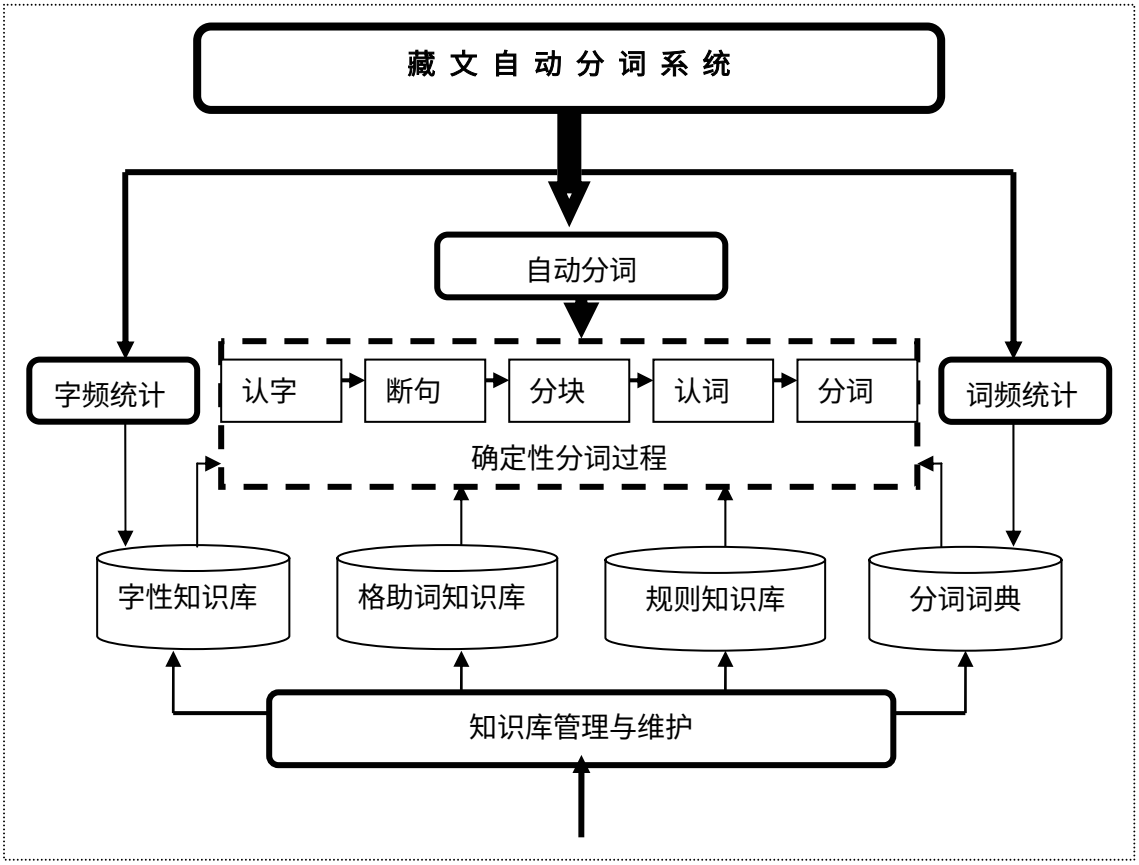


图 1 书面藏文自动分词系统结构图

2．2 主要功能模块

整个藏文自动分词系统从功能上可以分为自动分词、知识库管理与维护、字频统计、词频统计等四个主要模块。自动分词模块居于系统核心地位，称之为核心模块，其它三个模块在分词系统中相对地处于辅助地位，称之为辅助模块。在此，我们首先就三个辅助模块的主要功能作一简要介绍，核心模块的功能及其实现算法作为本文的主要任务留待后面的三节中再来作详细讨论。

---

[1] 格助词有狭义和广义之分，狭义上是专指传统语法“八格”理论中除第一格和第八格之外的其他六格所涉及的语法虚词；而广义上则包括传统语法中所讲到的所有的虚词。除非特别说明，本文谈及的格助词均指广义的格助词。同时，为了称说上的方便，除格助词以外的虚词我们暂称为接续词，格助词和接续助词统称为语法虚词。

[2] 接续特征的说明参见本文 3.4 及其注解。

2．2．1 知识库管理与维护

该模块的主要功能是维护和管理分词系统的四个知识库，即字性知识库、格助词知识库、接续特征规则知识库和分词词典的管理和维护。

2．2．2 字频统计

用于统计待切分语料中藏字的字数。

2．2．3 词频统计

可根据需要随时统计待切分语料中词的频度信息，为今后开展基于规则和统计相结合的藏文分词系统研究提供服务。

3、基于接续特征的知识库组织与实现

分词知识库的质量和规模是决定分词精度的关键因素。分词知识库的组织形式又与具体采用的分词策略息息相关，而且还对分词系统的通用性以及分词效率有着较大的影响。本分词系统采用的是多级分词策略，因此，针对分词过程中不同的处理层面对分词知识的不同需求建立了相应的知识库。分别包括主要与认字和断句一级相对应的字性知识库，主要与分块一级相对应的格助词知识库，主要与认词、分词一级相对应的分词词典和接续特征规则知识库。在系统设计过程中，我们尽可能采用了统一的格式和数据结构来组织不同类型的知识库，以便于程序设计和具体实现。

3．1 基于字接续特征的字性知识库组织

字性知识库包含了按藏文拼读规则所能拼写出的所有字以及特有的藏文标点符号等藏文字符 14400 余个。根据藏文分词和进一步句子分析的需要，主要对接续特征、构字部件、字性及其组成数量等藏字的基本属性进行了详细的描述。

3．2 基于句节接续特征的格助词知识库组织

藏语句子的表达主要借助格助词来完成，格助词的种类和所添接位置的正确与否直接关乎句子所表达的意思。书面藏文格助词有 82 个（包括变体）[2]，根据是否受后置字约束分为规则格助词和不规则格助词两大类。规则格助词主要分 7 类共 63 个，不规则格助词主要分 6 类共 19 个。对这些格助词，我们根据句节[1]接续关系和字接续特征建立了格助词知识库。

3．3 基于词接续特征的分词词典组织

词典的规模、质量和信息容量已成为衡量某种语言自然语言处理发展水平的关键指标之一。因此，学界历来对电子词典的建设都非常重视。目前本系统分词词典的词条规模约 10 万余条，每条词都标注了词的接续信息。结合藏文分词的实际需要和进一步开展藏文词性标注研究的需要，根据功能分类思想[3]，提出了信息处理用现代藏语词语的分类方案，并依据这一方案对现有电子词典中的词语进行了归类。

3．4 基于三级接续特征的规则知识库组织

从分词角度来看，藏文的特征包括字切分特征、词切分特征和句切分特征等三级接续特征[2]。在接续特征知识库中，各级接续特征知识采用产生式规则表示。同时，根据分词过程中的实际需求，对各类接续特征规则进行了综合归类，形成不同级别的规则，统一存储在一个知识库中。

---

[1] 句节的含义参见本文“4.1 分词策略”第 3 自然段。

[2]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。

4、系统分词策略和分词算法

4．1 分词策略

藏语是黏着性语言，传统藏文语法历来重视格助词及其接续特征的研究，在一定意义上，我们可以把传统藏文语法看作是由格助词及其接续特征规则构成的语法系统。这一语法系统的主要特点就是：各类名词性成分借助格助词及其接续特征规则构成句节进而由句节结合动词来组织句子。一般而言，藏语的句子又是以动词为中心来组织的，动词居于句子末尾，制约着全句的格局，决定着格助词的添接规则。人们之所以能够阅读并理解句子的含义，主要靠的就是这种词之间、短语之间的格助词及其接续特征规则以及由此构成的句节与句末动词之间特有的相互联系。一个句子如果在格助词的表述上出现了错误，我们就无法正确理解它的含义。进而言之，藏语句子的组织过程就是在词与词、短语与短语之间添加格助词并与句末动词有效地结合的过程，而藏语分词过程则相当于组织藏语句子的逆过程。因此，藏文分词的关键是如何结合藏语字、词、句各类形式特征来确定格助词及其接续特征规则的识别算法、结合分词过程的实际需求来有效利用各类资源并进而制定出切实可行的藏文分词策略。

根据藏文格助词的接续特征，我们把格助词及其接续特征规则的识别问题可转化为“能够”接续和“应该”接续两个问题。而这两个问题的解决必须从句子和字两个层面来入手，字层面要识别格助词的可接续性，即“能够”接续问题；句层面要确定格助词及其接续特征规则的动词制约性，即“应该”接续问题。“能够”接续问题可通过后置字与格助词接续信息的合一运算来实现，而“应该”接续问题可通过格助词信息与动词的一致性检查来实现。为此，要识别藏语句子中的格助词及其接续特征词并进而进行分词首先要解决认字和断句问题，这就需要字性知识库、接续特征规则知识库和格助词知识库的支持。

藏语句子的各个功能性成分主要是词和格助词及其接续特征词的结合体，同时还有一些则是短语或子句与格助词组成的连续结合体，我们统一称之为句节（或块）。由短语或子句组成的句节内词的切分必须借助词典和接续特征规则。句节内无法切分的“堆块”以及由属格格助词引起的“截断”问题在分析阶段需综合各类知识才能解决。据此，我们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”的多级切分策略。其基本处理流程如图 2 所示。

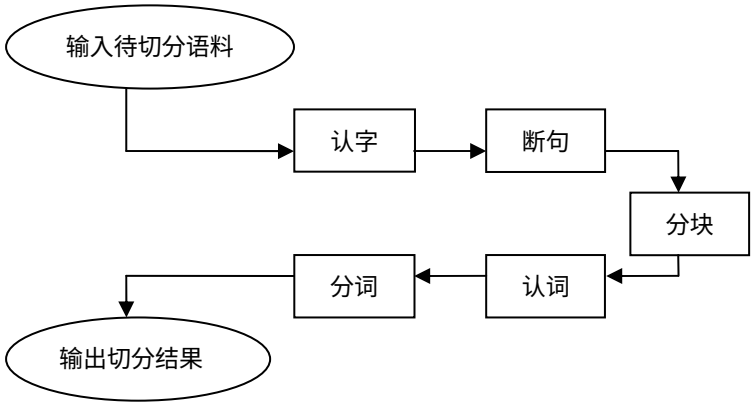


图 2 基本处理流程

4．2 BCCF 分词算法

分词算法是分词策略的形式化描述和具体实现，由以上分词策略可知，BCCF 算法（基于格助词和接续特征的分词算法）的主要特点是：综合利用书面藏文字、词和句的接续特征

进行的确定性分词。具体算法如下：

（1） 认字和断句

读入字符（直到句末或文件尾）；

if 字长<3

if 格助词

赋格助词属性；

else 赋字属性；

（2） 格助词识别和分块（直到句末）

if 无歧义格助词

加切分标记，继续（2）；

else if 规则格助词

取前一字的后置字，进行字接续合一运算；

if 满足合一条件

取其邻接或句末动字的格属性值，进行合一运算；

if 满足合一条件

加切分标记，继续（2）；

else 加歧义切分标记，继续（2）；

else 调用格接续特征规则处理；

（3） 认词（直到句末）

if 块长<2

加切分标记；

else 调用词典进行查找匹配并进行词接续特征检验；

（4） 分词（直到句末）

if 非句末动词、非格助词且该词串的后置字符合紧缩格的条件

取其邻接或句末动字的格属性值进行合一运算；

if 满足合一条件

切分并赋格助词属性；

else 不切分

if 块内单字串长>2

加堆块标记，但不切分；

if 属格助词且其前或后词长<2

加截断标记，但不切分；

输出切分结果。

本算法中，规则格助词的识别主要依据前接字的接续约束关系邻接或句末动字的格接续关系来统一处理，不规则格助词由于不受后置字约束，需要接续特征规则的支持。歧义切分标记、堆块标记和截断标记仅为进一步分析提供信息，切分系统中不作输出。

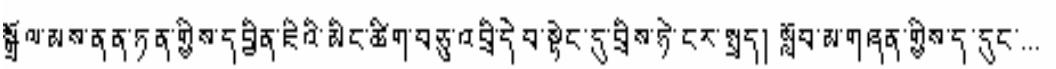
4．3 分词过程举例

以下主要结合 4.1 中提出的处理流程，通过一个分词实例来介绍 BCCF 分词算法的具体实现过程。

1) 认字、断句算法

认字和断句的基础是字性库、分字点、标点和句特征。

首先，待处理的原文片断如下：



(1)

[卓玛认真地将十个英语单词写在作业本上并交给了我，而其他同学还在……]
本句中，认字和断句过程的算法描述如下：

---

i. 读入字符串至自然标点符号为止；

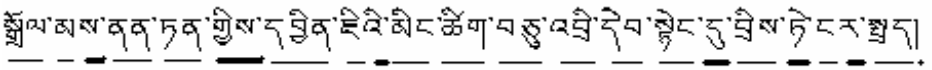
ii. 认字： C’+c 为非字，C’和 c 应分开，可以确定共有 19 个字、17 个分字点和 1 个标点；

iii. 确定字性特征：由字性知识库可以判定，除 c 外可能的格助词还有 b、d、e；可能的动字有 F、H、I、J 四个，由句特征信息可认定 J 为句末动字；

iv. 通过分字点确认 g 为自然句界而 D’为非句界；

v. 通过 J 的字性特征判定这是一个句子

通过断句和认字得到如下结果：



(2)

[卓玛认真地将十个英语单词写在作业本上并交给了我]

其实，这一步还无法识别 a、f 为格助词，在此之所以先用粗下划线与其他格助词一起标出来，完全是为了便于进一步的讨论。

2) 格助词识别和分块

格助词识别和分块的基础是字性库、格助词库和接续特征规则库。

格助词识别和分块的算法描述如下：

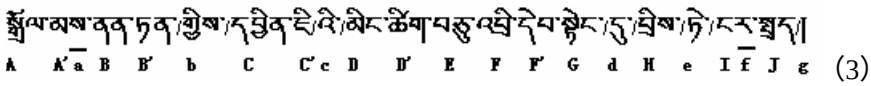
---

i. 识别无歧义格助词：由格助词库及其接续特征可确认 b、c、e 为无歧义格助词；

ii. 识别规则格助词：本句中没有不规则格助词，可能的规则格助词为 d，由其接续特征和前接字 G 的后置字、邻接动字 H 可判定 d 为格助词；

iii. 分块：通过以上两步得到的分块结果如（3）所示。

通过分块得到如下结果：



(3)

[卓玛认真地/将十个英语单词/写在/作业本上/并/交给了我/]

通过这一步无法判别的格助词只剩下a，f两个紧缩格。除此之外，本句中其他格助词基本确定。因而，可能的堆块有 5 个，即A-B’+b、C-C’+c、D-G+d、H+e、I-J+g。例（3）中，斜杠“/”用来表示块与块之间或词与词之间的分割符（下同）。

3) 认词

认词的基础是分词词典、接续特征规则库和格助词库。

认词过程的算法描述如下：

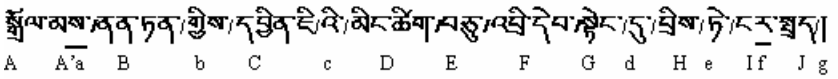
---

i. 认单字词：本句中只有 H 是单字词;

ii. 认多字词：由 A-B’>1,查词典得到 A、A’和 B，A、A’为堆块，B 为双字词；C-C’>1,查词典得到 C、C’为堆块；D-G>1,查词典得到 D、E、F、G；I-J>1,查词典得到 I、J 为堆块；

iii. 认词得到的结果如（4）所示。

通过这一步得到如下结果：


A A'a B b C c D E F G d H e If J g (4)

[卓玛/认真/地/将/十/个/英语/单词/写/在/作业本/上/并/交给了我/]

（4）中有 5 个单字词、4 个格助词和 4 个双字词。可能的动词有 3 个，即 H、I、J。除此之外，可能的截断错误有 1 个，可能的堆块错误有 3 个。

4) 分词

分词的基础是字性库、分词词典、接续特征规则库、格助词库。

分词过程的算法描述如下：

---

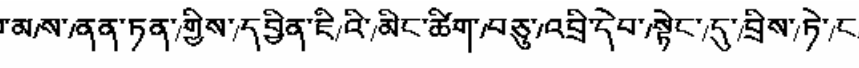
i. 紧缩格识别：本句中，a、f 及 H 的后置字有可能是紧缩格。由 d、e 接续特征规则和 H 的字性特征可判定 H 的后加字不是格助词；由 H 的字性特征、e 的接续特征可判定 a 是紧缩格；由 J 的字性特征和 e 的接续特征可判定 f 为紧缩格；

ii. 无堆块标记；

iii. 无截断标记；

iv. 输出切分结果。

通过以上各步处理，最后得到如下正确的分词结果（5）。

 (5)

[卓玛/认真/地/将/十/个/英语/单词/写/在/作业本/上/并/交/给/了/我/]

4．4 分块策略的再讨论

4．4．1 “ 分块 ” 策略与切分标记法

分块策略与汉语分词中曾经采用过的 “ 切分标志法 ” [4]在本质上是相近的。汉语自动分词研究中一般都认为 “ 切分标记法 ” 是一个毫无必要的技术，它增加了时空复杂性，却并没有提高分词精度。由于汉语 “ 重意合、轻形式 ” 的特点，实际上对汉语而言，“ 切分标记 ” 也确实没有标记歧义字段的任何信息。但这一相近的方法在不同的语言中所发挥的效能则完全不同，特别是对藏语这样的粘着性语言而言，有效地利用 “ 分块 ” 策略是非常有实际意义的。其理由主要有二：从分词本身来说，最好的选择是能用各类形式特征标记来完成切分。而藏语格助词是句子中词与词之间、句节与句节之间 “ 天然 ” 的形式特征标记，因而，首先利用格助词进行分块是最自然最方便的选择。从自然语言处理的目标来看，词的切分只是其中的一项基础性工作，在此基础上开展的句子分析和理解才是最终目标。而要分析和理解藏文句子，格助词的正确识别又是至关重要的。因此，在藏语中 “ 分块 ” 策略不但有利于分词工作本身，而且是进一步开展句子分析和理解的基础。

4．4．2 “ 分块 ” 策略与 “ 未登录词 ” 识别 通常分词系统采用的未登录词识别策略是在分词 “ 碎块 ” 中采用一定的算法查找并重新组合[5]。且不说未登录词是不是一定都包含在分词 “ 碎块 ” 中，就是假定所有的未登录词都分布在某一段分词 “ 碎块 ” 内的前提下，仍然面临着一个无法克服的技术难题：即分词系统事先无法确认组成未登录词的 “ 碎块 ” 到底是一些单字串还是几个双字串或是包含单、双字串在内的多字串。也就是说我们无法识别组成未登录词 “ 碎块 ” 的边界和大小。但在采用基于格助词和接续特征分词方法的情况下，如果存在某个无法切分的块我们就基本可以确定这个块可能就是一个未登录词，在知识尚不完备的分词阶段，我们采取加标记但不切分的 “ 谨慎 ” 策略。这样，不但解决了未登录词的识别

问题而且很自然地确定了其相应的边界和大小，这无疑对进一步开展句子分析是非常有积极意义的（相关研究工作另撰文汇报）。

5、实验结果和结论

本系统采用的基于格助词和接续特征的藏文分词方法是一种利用藏文字、词、句各类接续特征的分词方法，其主要目的在于提高自动分词的精度。通过对各类语料的对比测试表明，系统切分精度均达到了 96%以上。同时，实验结果还表明，本系统对不同领域的藏文语料表现出较强的适应性，因而说明了系统具有较强的通用性。

藏文自动分词是藏语信息处理中的基础性课题，由于自然语言固有的复杂性，加之本系统主要又是基于规则的系统，对藏语句子中的诸如堆块、词截断等错误的处理能力方面还有待改进。在此基础上，积极进一步引入统计技术，尝试规则与统计相结合的分词技术研究进而开展藏文分词、词性标注和句子分析的一体化分析技术研究，是我们下一步的努力方向。

参 考 文 献

[1] 陈玉忠、李保利、俞士汶、兰措吉，基于格助词和接续特征的书面藏文分词方案，语言文字应用，2003 年第 1 期。

[2] 才旦夏茸，ཨོད་གངས་ཅན་གྱི་སྒྲ་རིག་ལའི་བསྟན་བཅོས་ཤེན་འགའ་ཞིག་བསྡུས་བཞུགས་སོ། （藏文文法详解），青海民族出版社，1988。

[3] 朱德熙，语法讲义，北京：商务印书馆，1999。

[4] 刘挺、吴岩、王开铸，串频统计和词形匹配相结合的汉语自动分词系统，中文信息学报，1998 年第 1 期。

[5] 陈小荷，自动分词中未登录词问题的一揽子解决方案，语言文字应用，1999 年第 3 期。