基于法律文本的藏语句子边界识别*

赵维纳 [1,2,3]，刘汇丹 [3,4]，于新 [3,4]，吴健 [3]，张普 [1]
1. 北京语言大学，北京 100083；2. 青海师范大学，西宁 810008；
3. 中国科学院软件研究所，北京 100190；4. 中国科学院研究生院，北京 100190
E-mail:zhaoweina1999@yahoo.com.cn, huidan@iscas.ac.cn,
yuxin08@iscas.ac.cn ,wujian@iscas.ac.cn,zhangpu@blcu.edu.cn

摘要：由于传统藏语书写中标点符号的特殊性，在藏语句子中如何找到正确的句子边界，是正确识别句子的首要问题。本文通过对藏语法律文本语料的观察，利用规则提出了一个藏语句子边界的识别算法。同时对藏语法律文本的句式特点进行初步的分析和探讨。

关键词：断句;藏文句子边界识别;藏文信息处理;中文信息处理

The Tibetan Sentence Boundary Identification Based on Legal Texts

Zhao Weina[1,2,3],Liu Huidan[3,4],Yu Xin[3,4],Wu Jian[3],Zhang Pu[1]
1. Beijing Language and Culture University, Beijing, 100083;
2. Qinghai Normal University, Xining, 810008;
3. Institute of Software, Chinese Academy of Sciences, Beijing, 100190;
4. Graduate University of the Chinese Academy of Sciences, Beijing, 100190
E-mail: zhaoweina1999@yahoo.com.cn, huidan@iscas.ac.cn,
yuxin08@iscas.ac.cn ,wujian@iscas.ac.cn,zhangpu@blcu.edu.cn

Abstract: As the specificity of the Tibetan punctuation marks, to find the correct sentence boundary has become an urgent task of Tibetan sentence identification. We propose an algorithm of Tibetan sentence boundary identification with rules which based on our research of legal texts. Meanwhile, we analysis and discuss features of Tibetan sentences.

Key words: sentence boundary identification; Tibetan sentence boundary identification; Tibetan information processing; Chinese information processing

1 引言

藏语是一门古老的语言，据记载从公元7世纪创造起至今已有1400多年的历史。藏语自身有自己的文字、语音和语法系统。在藏语书写中自身具有独特的标点符号体系，发展至今在现代藏语书面语中，仍然有所使用。

藏语标点符号是一套未臻完备的符号系统，主要表现为变体形式较多、意义含混、功能不确定。这种特征突出地表现在句子的断句方面，为此，藏语文本处理中要考虑句子边界识别(又称

---

*本文承中国科学院西部行动计划高新技术项目（KGCX2-YW-512）的资助。

分句)问题。

藏语中有自己特定的一套标点符号系统，不同于汉语或英语中的常见的“。”句号、“，”逗号等标点符号。藏语中涉及分句的标点符号主要有[1]：

音节点“་”：藏语称之为（ཚེག），由于藏文是以音节为单位的拼音文字，从左向右横向书写，音节点用以划分音节与音节的间隔；

楔形号“།”：藏语称之为（ཤད།），用以表示藏文的词、句子的停顿或结束。楔形号中又可以分为单楔形号“།”（ཚེག་ཤད།）亦称“单垂符”或“垂符”、双楔形号“།།”（ཉིས་ཤད།）、四楔形号“།།།།”（བཞི་ཤད།）；

起始号“༄༅།།”：藏语称之为云头符（མགོ་ཡིག།），多用于文章、书籍篇首，表示文字由此开始；

其次还有敬重号和着重号，多用于古籍和经卷当中，现代书面语当中较少。例如：

༄༅།། ཀྲུང་ཧྭ་མི་དམངས་ སྤྱི་མཐུན་རྒྱལ་ཁབ ཀྱི་ཀྲུའུ་ཞིའི་བཀན

中华人民共和国法律汇编

在这些标点符号中单垂符在藏文句子中主要用于句末，也用于词或者短语之后。从功能上分析相当于汉语标点符号的顿号、逗号、分号、句号、叹号以及问号。例如：

དོན་ཚན་བཞི་པ་རྫོང་རིམ་པ་ཡན་གྱི་མི་དམངས་སྲིད་གཞུང་གི་ཁོར་ཡུག་སྲུང་སྐྱོང་སྲིད་འཛིན་གཙོ་སྐྱོང་སྡེ་ཁག་གིས་རླུང་ཁམས་ཆེན་པོའི་སྦགས་བཙོག་འགོག་སེལ་བྱ་རྒྱུར་གཅིག་གྱུར་གྱིས་ལྟ་སྐུལ་དང་དོ་དམ་བྱ་རྒྱུ། སྤྱི་བདེ་དང་། འཁྲིམ་འགྲུལ། ལྕགས་ལམ། ཉ་ལས་བཅས་ཀྱི་དོ་དམ་སྡེ་ཁག་རིམ་པ་ཁག་གིས་སོ་སོའི་འགན་འཁྲི་ལ་གཞིགས་ཏེ་འཕྲུལ་སྐྱུལ་འཁོར་ལོ་དང་གྲུ་གཟིངས་ནས་རླུང་ཁམས་ཆེན་པོ་སྦགས་རྫུལ་ལ་ལྟ་སྐུལ་དང་དོ་དམ་བྱ་རྒྱུ།

第四条 县级以上人民政府环境保护行政主管部门对大气污染防治实施统一监督管理。各级公安、交通、铁道、渔业管理部门根据各自的职责，对机动车船污染大气实施监督管理。

从上述例子的汉语句子中我们可以清晰的从句号、逗号、顿号的位置直接识别出句子的边界和标点符号的功能，但是对应的藏语句子中，一共出现了6个相同的单垂符，无法直接从单垂符的形式上区分到底在句子中所起的作用和句子的边界。

由于藏语标点符号的特殊性，导致藏语句子里句终符号的不确定性，机器在识别整句上存有难度。江荻首次对藏语的疑问句进行识别，并提出了有关藏文句子边界的问题，指出“藏语文的自然语言处理不仅有分词的问题还有分句的问题”[2]。相同的问题在英文文[3]-[11]本和维语[12]文本中存在。

2 藏语句子语尾分析

藏语的句子语序结构不同于汉语或英语的语序结构，属于SOV型，即{主语+宾语+谓语}的

句子语序，其中谓语部分中包括整个句子中的核心动词。所以在一个完整的藏语句子中，核心动词的句法位置始终位于句子的末端结尾部分。因此藏语句子中谓语部分的末端应当是整个句子的煞尾结点。但是单独对于谓语部分做分析时，其谓语结构又有所多样化，单独以动词煞尾的句子不多见。一般在句子的谓语部分中核心动词后边总是附加包含有一些其他成分，这些成分可统称为动词的语尾，其谓语的语序格式为：{（谓语动词（+状语补语）(+助动词[情态和趋向]）（+体貌-示证标记）（语气词）}[13]。

相对于其他题材的文本，法律文本的句式结构相对比较规范和整齐，其句式结构也相对单一。通过对《中华人民共和国法律汇编 20 00 藏》观察分析，在该类文本中主要由以下 18 种形式的语尾形式。

དགོས།	མི་དགོས།	ཆོག	མི་ཆོག	རུང་།
མི་རུང་།	ཀྱི།	རྒྱུར།	རེད།	ཡིན།
ཡོད།	འདུས།	ཟེར།	མི་ཚུད།	མེད།
སསོ།	སྐོར།	ལས་ཆེ།		

①助动词结尾

以助动词直接作为句子的结尾是法律文本中常见的一类形式，可分为能愿助动词和事态助动词。能愿助动词用于表达一定的情态或愿望。

- 情态助动词

在法律文本中主要出现有“应该、需要；可以；适合”三种意义的情态助动词，以及这三种词的否定形式，详见下表及部分代表例句。

དགོས།	མི་དགོས།	ཆོག	མི་ཆོག	རུང་།	མི་རུང་།
应该、需要	否定形式	可以、行了	否定形式	可以、适合	否定形式

例句：

句尾དགོས།（应该、需要）

ཁྲིམས་འཛུགས་སྐབས་ཁྲིམས་བཀོད་དབང་ཚད་དང་གོ་རིམ་ལྟར་རྒྱལ་ཁབ་སྤྱི་ཡོངས་ཀྱི་ཁེ་ཕན་གཞིར་བཟུང་སྟེ་སྤྱི་ཚོགས་རིང་ལུགས་ཀྱི་ཁྲིམས་ལུགས་ཀྱི་གཅིག་གྱུར་དང་གཟི་ཛམ་སྲུང་སྐྱོང་བྱ་དགོས།

立法应当依照法定的权限和程序，从国家整体利益出发，维护社会主义法制的统一和尊严。

句尾ཆོག　（可以、行了）

ཆེད་དོན་ཨུ་ཡོན་ལྷན་ཁང་གིས་གྲོས་ཞིབ་བྱེད་དུས་གྲོས་གཞི་འདོན་མཁན་གྲོས་ཚོགས་ལ་ཟུར་ཞུགས་དང་བསམ་འཆར་འདོན་པར་གདན་འདྲེན་ཞུས་ཆོག

专门委员会审议的时候，可以邀请提案人列席会议，发表意见。

- 时态助动词

除了在使用上述情态助动词外，还用到两个时态助动词རྒྱུ།表示“将要”，བྱུང།表示过去的时间，类似于“了”。其中表示“将要”རྒྱུ།占到了大多数，这也体现了法律文本的一个特点。例句：

ལས་ཁུངས་ཁག་དང་ཚ་འཛུགས་ཁག་སྤྱི་དམངས་བཅས་ཀྱིས་བཏོན་པའི་བསམ་འཆར་རྒྱུན་ལས་ཨུ་ཡོན་ལྷན་ཁང་གི་ལས་ཁུངས་ལ་སྤྲོད་རྒྱུ།

各机关、组织和公民提出的意见送常务委员会工作机构。

②直接以动词结尾

藏语句子中直接以动词作为句子结尾的句尾并不多见（ཡིན།（是）、རེད།（是）、ཡོད།（有）等存在动词或判断动词除外），对该法律文本观察的到以下：འཐུས།（适合、恰当）、ཟེར།(说)、ཡིན།（是）、ཡོད།（有）、རེད།（是）、མི་ཚུད།（除外）六种直接以动词结尾的句子。例句：

ལས་ཁུངས་སུ་ཀུའུ་ཀྲང་དང་། ཁེའོ་ཀྲང་། ཁྲུའུ་ཀྲང་བཅས་དང་རིམ་མཚུངས་འགོ་ཁྲིད་གོ་གནས་འཛིན་ཡུན་ལོ་བཞི་ཡིན་པ་དང་། ཅུས་ཀྲང་དང་པུའུ་ཀྲང་དང་རིམ་མཚུངས་འགོ་ཁྲིད་གོ་གནས་འཛིན་ཡུན་ལོ་ལྔ་ཡིན་པ། ཡིན་ནའང་ཆེད་ལས་ཀྱི་རང་བཞིན་ཆེས་ཆེར་ལྡན་ཞིང་ལས་ཀར་དགོས་མཁོ་དམིགས་བསལ་ཆེ་བའི་མི་ཉུང་ཤས་ཤིག་དེའི་ཁོངས་སུ་མི་ཚུད།

机关担任股长、科长、处长及相当领导职务的，四年；担任局长、部长及相当领导职务的，五年；但是少数专业性强和工作特别需要的除外。

③直接以否定词结尾

藏语的否定词一般有四种形式མ་མི་མིན་མེད་。其中མ་和མི་是表否定语气的副词，置于被否定的动词之前，མིན་和མེད་是表示否定语气的动词，置于否定词语的后面[14]。在该文本中只出现མེད་作为结尾的句子，例如：

ཐོན་རྫས་དེར་འཛོམས་དགོས་པའི་བེད་སྤྱོད་བྱེད་ནུས་འཛོམ་དགོས། ཡིན་ནའང་ཐོན་རྫས་དེའི་སྤྱོད་ནུས་ཐད་སྐྱོན་ཆ་ཅི་ཡོད་གསལ་བཤད་བྱས་རིགས་དེའི་ཁོངས་སུ་གཏོགས་མེད།

具备产品应当具备的使用性能，但是，对产品存在使用性能的瑕疵作出说明的除外。

④以终结词སོ།结尾

藏语终结词是指用于句末表示句子结束的词语，一共有 11 种不同形式的变体（གོངོདོནོབོམོའོརོལོསོཏོ），但作用一致在使用时主要依据前一音节的后加字配合使用。在文本中只出现一种形式“བཅས་སོ།”。句终词一般在早期的藏文文本中，现代文本中通常将终结词语尾用于多重复句的结尾。例如：

ཐོན་སྐྱེད་བྱེད་མཁན་གྱིས་གཤམ་གསལ་གྱི་གནས་ཚུལ་ལས་རིགས་གཅིག་ཡོད་པའི་ར་སྤྲོད་བྱེད་ཐུབ་ན་སྐྱིན་ཚབ་སྤྲོད་འགན་འཁུར་མི་དགོས།

གཅིག། ཐོན་རྫས་དེ་འགྲོ་རྒྱུག་བྱེད་པར་བཏང་མེད་པའི་རིགས་དང་།

གཉིས། ཐོན་རྫས་འགྲོ་རྒྱུག་བྱེད་པར་གཏོང་སྐབས་གནོད་སྐྱོན་བཟོ་སྲིད་པའི་སྐྱོན་ཆ་མེད་པའི་རིགས།

གསུམ། ཐོན་རྫས་འགྲོ་རྒྱུག་བྱེད་པར་གཏོང་སྐབས་ཚན་རིག་ལག་རྩལ་གྱི་ཆུ་ཚད་ཀྱིས་སྐྱོན་ཆ་ཡོད་མེད་གསལ་ལ་ཕེབས་བྱེད་མི་ཐུབ་པའི་རིགས་བཅས་སོ།

生产者能够证明有下列情形之一的，不承担赔偿责任：（一）未将产品投入流通的；（二）产品投入流通时，引起损害的缺陷尚不存在的；（三）将产品投入流通时的科学技术水平尚不能发现缺陷的存在的。

⑤其他形式结尾

除了上述几种形式外，还有以下两种སྐོར།和ལས་ཆེ།，其中སྐོར།通常用在标题语的后边意思为“这一类”。这是一个名词，在法律文本中大量出现在标题句子中，所以也可认为是一种句尾形式。例如：

ས་བཅད་དང་པོ།ཁྲིམས་འཛུགས་ཀྱི་དབང་ཚད་སྐོར།

第一节立法权限

而ལས་ཆེ།（较大、较高）是一种以形容词结尾的句式，在藏语的比较句中通常是将表示比较的形容词放在句末。在文本中出现了共 4 次，但是对于句子边界识别来说，比较句的边界较为特殊应进行单独的识别。

ས་གནས་ཀྱི་ངོ་བོ་ལྡན་པའི་ཁྲིམས་སྲོལ་གྱི་ནུས་པ་ནི་རིམ་པ་དེ་ག་དང་རིམ་པ་འོག་མའི་ས་གནས་སྲིད་གཞུང་གི་སྒྲིག་སྲོལ་ལས་ཆེ།

地方性法规的效力高于本级和下级地方政府规章。

3 边界识别分析

在分析清楚了主要句型的结尾形式后，可以依照上述 18 个句尾标记词进行句子的逆向匹

配，可直接识别出句子边界，达到断句的效果。这里需要指出的是，由于法律文本的句式结构在使用上相对结构形式单一，所以不需要额外考虑句尾词语的歧义现象，以及句子的谓语动词的搭配情况，故可直接做句尾的匹配。另外，在句子边界识别过程中由于藏语连词的特殊性，所以对于复句的识别未进行额外的处理。句子边界识别算法步骤如下：

1) 对藏文文本进行预处理，去掉冗余信息，得到待处理字符串(S)；
2) 如果 S 不为空，从左到右扫描 S，读入当前字符(char)到候选句子数组($W_{[i]}$),将指针(pointer)前移；
3) 判断 char 是否为"།"或者"ག།+一个空格"；
4) 如果 3)为真，查断句语尾表，利用最大逆向匹配方法，判断数组 $W_{[i]}$是否含有句子语尾串；
5) 如果 $W_{[i]}$含有句子语尾串，判断数组 $W_{[i]}$的含有音节的数量是否大于 5；
6) 如果 5)为真，将 $W_{[i]}$作为一个句子输出，S=S- $W_{[i]}$，返回步骤 2)；
7) 如果 5)为假，返回步骤 2)；
8) 如果 S 为空，输出 $W_{[i]}$，程序结束。

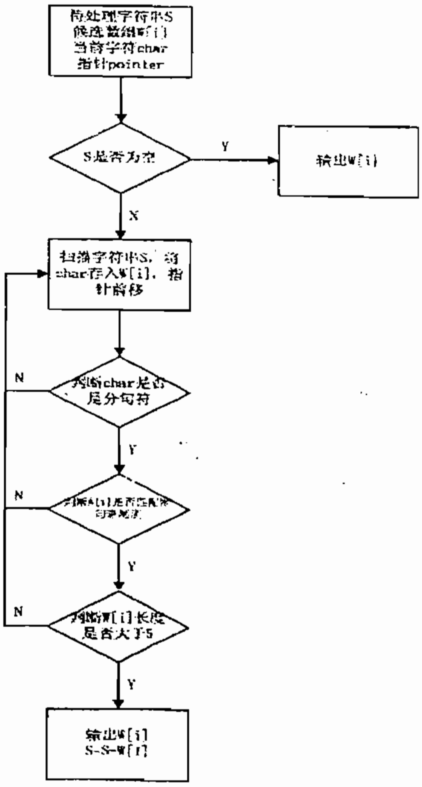

图 1 算法流程图

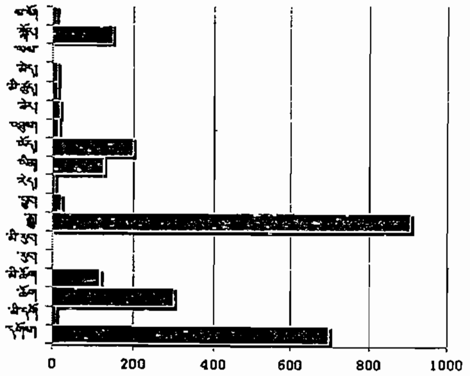

图 2 18 种句尾数量图

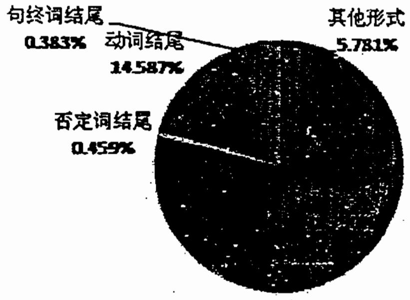

图 3 句子类别比例

对于整个《中华人民共和国法律汇编 20 00 藏》的句子边界识别后，识别句子数共 2612 句。各类句子数量见图 2 所示。

对于上述 18 种句尾的 2612 个句子按照句式结构可以分为助动词结尾句、动词结尾句、否

定词结尾句、句终词结尾和其他形式结尾句，在语料中的比例如图 3 所示。

但是在识别过程中，应当对于句子的长度有所限制，有些情况下有些短语中包含有句尾词，并不是真正意义上的一个完整句子的结尾，例如：

འོན་ཀྱང་ཉེས་གསོག་དང་ཉེས་ཆད་གཅོད་རྒྱུ་དང་། སྤྱི་དམངས་ཀྱི་ཆབ་སྲིད་ཁེ་དབང་བཙན་འཕྲོག་དང་མི་ལུས་ཀྱི་རང་དབང་ལ་བཀག་སྡོམ་བྱེད་པར་བཙན་

ཤེད་ཀྱི་བྱེད་ཐབས་སྤྱོད་རྒྱུ། ཆད་པ་གཅོད་རྒྱུ། ཁྲིམས་འཛིན་ལམ་ལུགས་ལ་སོགས་པའི་དོན་ཚན་འདིའི་ནང་མི་ཚུད།

但是有关犯罪和刑罚、对公民政治权利的剥夺和限制人身自由的强制措施和处罚、司法制度等事项除外。

该句中含有 2 个རྒྱུ།的句尾，如果直接做断句处理会在“ཆད་པ་གཅོད་རྒྱུ།”的后面断开，所以在句子长度上有所限制后，会避免类似情形的发生。

4　结论

由于藏语中标点符号的特殊性，使得藏语句子的边界难以直接做出准确的判断，但是通过对于句式结构的分析可以在一定程度上判断出部分句子的边界。因而得出对于藏语句子虽然没有形式上的类似英语或汉语中的句号等句终符号标记，但是可通过句尾信息的判断可以到达分句的目的。本文通过对法律文本的观察，得出 18 种句尾形式从而达到分句的效果。

参考文献：

[1]　胡书津 《简明藏文文法》云南民族出版社

[2]　江荻 藏语的疑问句与句界识别

[3] Palmer D D ,Hearst M A. Adaptiv e multiling ual sen tence bou dary dis ambiguation[J]. Computatonal Linguistics, 1997,23(3):241-267

[4] Reynar J C, Ratnaparkhi A. A maximum entropy approach to indentifying sentence boudaries[C]. In: Proceedings of the Fifth ACL Conference on Applied Natural Language Processing(ANLP'97), Washington,D.C.,1997

[5] Mikheev A. Tagging sentence boundaries[C]. In NACL'2000ACL,2000,264-271

[6] Dan Gillick. Sentence Boundary Detection and the Problem with the U.S., Proceedings of NAACL HLT 2009: Short Papers, pages 241–244

[7] J effrey C Reyaar and Adwait Ratnaparkhi，A Maximum Entropy Approach to Identifying Sentence Boundaries．1997

[8] Andrei Mikheev，Taggong Sentence Boundaries-Proc of NAACU2000

[9] 于中华,张容,唐常杰,张天庆 基于前后文词形特征的生物医学文献句子边界识别[J] ,小型微型计算机系统 2006

[10] 朱莉，孟遥，赵铁军，李生 英语句子边界的识别[C]，全国机器翻译研讨会 2002

[11] 阿比达・吾买尔，吐尔根・依布拉音 维吾尔语句子边界识别算法的设计与实现[J] 新疆大学学报 2008

[12] 江荻 《中国民族语言工程研究新进展》社会科学文献出版社

[13] 格桑居冕 《实用藏文文法》四川民族出版社