基于音节的藏语功能组块边界识别

王天航[1]，史树敏[1,2]，龙从军[3]，黄河燕[1,2]
(1.北京理工大学计算机学院，北京 100081；
2.北京市海量语言信息处理与云计算应用工程技术研究中心，北京 100081；
3. 中国社会科学院民族学与人类学研究所，北京 100081）
E-mail：{hbcdwth, bjssm}@bit.edu.cn, longcj@cass.org.cn, hhy63@bit.edu.cn

摘要：藏语句法功能组块分析旨在识别出藏语句子的句法成分。本文在藏语句法功能组块描述体系基础上，提出了一种不经过分词和标注，直接以音节为单位利用条件随机场(Conditional Random Fields，CRF)对句子功能组块边界进行识别的方法。根据藏语的实际特点，首先通过文本预处理，识别和提取出黏写形式和非黏写形式构成的句法标记作为句法功能块边界的识别特征，然后采用CRF模型进行句法功能块边界识别。在真实藏语语料（含46,783个音节）上开展实验，准确率、召回率与F值分别达到75.70%、82.54%与79.12%，实验结果表明该方法在小规模未标注语料上可以达到较为理想的组块识别效果，可以为后续机器翻译等自然语言处理应用提供基础支持。

关键字：藏语句法功能组块；组块边界识别；音节；句法标记；CRF；机器翻译

Tibetan Functional Chunks Boundary Recognition Based on Syllables

WANG Tianhang[1], SHI Shumin[1,2], LONG Congjun[3], HUANG Heyan[1,2]
(1.School of Computer Science & Technology Beijing Institute of Technology, Beijing 100081, China;
2.Beijing Engineering Research Center of High Volume Language Information processing & Cloud Computing Applications, Beijing 100081, China;
3.Institute of Ethnology & Anthropology Chinese Academy of Social Sciences, Beijing 100081, China)
E-mail:{hbcdwth, bjssm}@bit.edu.cn, longcj@cass.org.cn, hhy63@bit.edu.cn

Abstract: *Tibetan syntactic functional chunk parsing is aimed at identifying syntactic constituent in Tibetan sentences. In this paper, based on the Tibetan syntactic functional chunk description system, the author proposed a method which puts syllables in group instead of word segmentation and labeling and uses the Conditional Random Fields, CRF to identify the functional chunk boundary of a sentence. According to the actual characteristics of the Tibetan language, the paper firstly identifies and extracts, through the text pretreatment, the syntactic markers which are composed of the Sticky written form and the non-Sticky written form as identification characteristics of syntactic functional chunk boundary. And then it identifies the syntactic functional chunk boundary through CRF. Experiments have been made on 46783 syllables of Tibetan language corpora, and the precision, recall rate and F value respectively achieves 75.70%, 82.54% and 79.12%. The experiment results show that the proposed method is effective which can provide infrastructural support for machine translation and other natural language processing applications.*

Keywords: *Tibetan syntactic functional chunk; chunk boundary recognition; syllable; syntactic marker; CRF; machine translation*

基金项目：国家自然科学基金项目（61201352, 61132009）；国家重点基础研究发展规划(973)（2013CB329303）；北京理工大学基础研究基金（20130742010）

1 引言

句法功能组块分析是组块分析的一个重要组成部分。研究它的目的是标注出构成句子的核心句法成分块，通过自顶向下地进行句子拆分，从而获取句子中的基本信息单元，以显示句子在小句层面上的基本结构及骨架。在机器翻译中，以句法成分块为单位分析小句，可以有效地减小句子被过渡切分而导致词语重新组合时的难度；还有利于减少译文生成中的紊乱现象，在初步预处理基础上进行句法功能块的识别将对基于规则的机器翻译系统有十分重要的意义。

英语、汉语对组块的研究成果较多，这些成果可以为藏语功能组块的边界识别提供良好的借鉴。但以往功能组块研究基本上是在分词、词性标注的基础上进行的，一方面分词和标注的错误会直接影响句法功能块识别的正确性；另一方面也增加了句法功能块分析的时间和空间开销。从藏语目前的研究来看，分词和标注的准确率还有待提高，可以实用的词法分析软件较少。根据藏语具有较为丰富的形式句法标记这一特点，如果能够探索在原始语料基础上跨过分词和标注，直接进行句法功能块识别，也许会寻求到一条新的思路。

本文在藏语句法功能组块描述体系基础上，尝试不经过分词和标注，直接以音节为单位，利用条件随机场对句子的功能组块边界进行识别。在实验过程中，我们选择了句法标记作为特征加入到识别模型中，经过实验证明，在小规模未标注语料上可以达到较为理想的组块识别效果。

2 藏语句法标记

藏语是一种拼音文字，有 30 个辅音字母和 4 个元音字母，由这些字母组成音节，进一步构成词语，音节之间以“·”分开，与汉语类似，藏语的词语之间没有分隔符。藏语的句法标记比较丰富，这里所说的句法标记是指在现代藏语中，一些形式标记（包括格标记和助词标记）把句子分割成功能不同的句法组块，如表示时间、地点的状语的组块之后可能有处所格标记，主语组块之后可能有施事格标记，对象宾语组块之后可能有对象格标记等，这些标记是基于音节粒度的组块识别的基础。但是有一部分格和助词标记由于文字书写原因导致两个音节缩写为一个音节，称之为黏写形式[Long,2013]。为了能充分利用格及助词标记，我们不但需要独立音节的格及助词标记，也需要把那些构成黏写音节中的格及助词标记分离出来，这些标记共同构成句法功能组块边界的识别特征。

2.1 藏语黏写形式标记

藏语黏写形式有以下构成形式：(1)词+ས(-s)（施格/工具格标记）、(2)词+འི(-vi)（属格标记）、(3)词+ར(-r)（与格/位格标记）、(4)词+འང/འམ(vang/vam)（连词）、(5)词+ོ(-vo)（句终词）。但是各种黏写形式对于句法功能块识别的作用不均等，其中 ས(-s)和 ར(-r)在文本中频率高，对句法功能块识别最重要，འི(-vi)基本上属于块内标记，འང/འམ(vang/vam)和 ོ(-vo)有可能是块的边界，但是出现频率并不高。黏写形式的识别通常有三种处理方法：黏写形式与分词同步进行[刘汇丹,2012；Kang,2013；康才畯,2014]；黏写形式单独切分，又包括两种策略，一种只是识别切分黏写形式而对具体黏写形式不进行区分，则设计两个标签 N 和 Y，N 表示非黏写音节，表示黏写音节[李亚超,2013]；另一种是不但识别出黏写形式而且标记出黏写形式的类型，这样处理有利于后续组块边界和组块类型识别，本文即采用该种策略，设计的标签如下：

表 1 黏写形式类型

类型	例子	标签
音节+ས(-s)（施格/工具格标记）	བཏང་“我做”	S
音节+འི(-vi)（属格标记）	བའི་“我的”	V

音节+ར(-r)（与格/位格标记）	ངར་“对我”	R
音节+འང/འམ(vang/vam)（连词）	ངའང་“我也”	C

2.2 藏语非黏写形式标记

非黏写形式主要指独立音节的格标记和助词标记，格标记包括施事格、使动格、工具格 གིས、གྱིས、ཡིས、 ཀྱིས；la don 变体 སུ、རུ、ལ、ཏུ、དུ、ན 构成的时间格、处所格、对象格、领有格、向格；比较格 ལས；从格 ནས；伴随格 དང་ 等。助词包括比拟助词、停顿助词、枚举助词、方式助词、结果助词、目的助词等。这些标记可以根据功能不同设计不同的标注标签，在识别标记的同时确定组块的功能类型。但本文旨在边界识别，因此统一以一个标签 M 标注。

3 藏语功能组块

3.1 藏语功能组块体系

本文所使用的功能组块定义遵循文献[李琳，2013]的描述性定义，即主语块、谓语块、宾语块、状语块、补语块以及为了处理方便而增设的句法标记块。

3.2 藏语功能组块标注集

文章采用 BIE 标记集来标记功能组块，使得功能组块边界识别问题转化为一个序列标注问题。其中，功能组块起始位置标记为 B，内部位置标记为 I，结束位置标记为 E，功能组块之外的标点统一标记为 B。

以“ཁོང་གིས་ང་ལ་དེབ་དེ་འཕྲིར་འདུག”（他带给我那本书，khong gis nga la deb de vkhyer vdug.）为例，首先对该句进行句法标记识别，利用该标记集对其进行标记的中间结果为：

[ཁོང་གིས་][ང་ལ་][དེབ་དེ་][འཕྲིར་འདུག]进一步处理后得到的标注结果见图 1。

Eg1:ཁོང་/Bགིས་/E(M)ང་/Bལ་/E(M)དེབ་/Bདེ་/Eའཕྲིར་/Bའདུག/E

Latin:khong/B gis/E(M) nga/B la/E(M) deb/B de/E vkhyer/B vdug/E.

例 1:他带给我那本书。

图 1 藏语功能组块边界识别标注实例

4 基于条件随机场的藏语功能组块边界识别

4.1 条件随机场模型

条件随机场(Conditional Random Fields，CRF)是 John Lafferty 等人于 2001 年提出的一种基于统计的序列标注和分类模型，在本文中简单介绍 CRF 模型，详细信息见参考文献[John，2001]，它是基于无向图的条件概率模型，在给定需要标记的观察序列下，计算整个标记序列的联合概率，从而找出最优的标注结果。条件随机场具有表达长距离依赖性和交叠性特征的能力，能够较好地解决标注(分类)偏置等问题，并求得全局的最优解。对于给定的输入观察序列 x=x1x2…xn，其中每一个 xi 为一个词语，x 在组块分析中就是输入的词语序列，定义 y=y1y2…yn 为输出状态序列，在组块分析中的即是待标注的组块标记。对于给定参数以 Λ=λ1λ2…λk 的线性链 CRF，输入序列 X 的状态序列条件概率为:

$$P_\Lambda(y|x) = \frac{1}{z(x)} exp(\sum_{i=1}^{n} \sum_k \lambda_k f_k(y_{i-1}, y_i, x, t)) \qquad (1)$$

其中，$f_k(y_{i-1}, y_i, x, t)$表示一个特征函数，Z(x)为归一化函数，λ_k是与f_k相关的权重参数，可在训练中得到。

实验使用 TakuKudo 开发的开源 CRF++[1]来实现基于 CRF 句法标记处理与功能组块边界识别的序列标记任务。

4.2 文本预处理

在对功能组块边界识别之前，首先对原始文本进行黏写形式和非黏写形式句法标记识别预处理。实验采用 CRF 对句法标记进行识别，以音节本身以及音阶上下文特征信息作为特征模板信息，如表 2 所示。

表 2 原子特征模板

编号	模板	编号	模板
1	CurSyllable	4	Syllable+1
2	Syllable-2	5	Syllable+2
3	Syllable-1		

其中，Syllable 表示音节，试验窗口取 5，上表中+/-表示当前音节后/前所对应的音节。对以上的单一特征进行复合，得到以下复合特征模板，如表 3 所示。

表 3 复合特征模板

编号	模板
6	CurSyllable,Syllable-1
7	CurSyllable,Syllable+1
8	Syllable-1,Syllable+1

4.3 基于 CRF 的藏语功能组块边界识别

对原始语料进行句法标记识别后，在基于音节标注的藏语功能组块边界识别方法中，我们利用 3.2 所示的标注方法对语料中每个音节进行标注。而在功能组块边界识别中，本文定义如下原子模板，如表 4 所示。

表 4 原子特征模板

编号	模板	编号	模板
1	CurSyllable	6	CurCase
2	Syllable-2	7	Case-2
3	Syllable-1	8	Case-1
4	Syllable+1	9	Case+1
5	Syllable+2	10	Case+2

其中，Syllable 表示同前所示表示音节，Case 表示文本预处理的识别结果，当特征函数取特定值时，则该模板被实例化。

Eg3:[ང/][ས་/][གསར་འགྱུར་/བཤད་/མཁན་/དེ་/][ངོ་ཤེས་/ཀྱི་ཡོད་/]།/

Latin: ngas gsar vgyur bshad mkhan de ngo shes kyi yod.

例 3：我认识那个播音员。

图 2 原子模板特征选择示例

其中，以词“གསར་འགྱུར་”(新闻，gsarvgyur)中“གསར་”(新，gsar)为当前音节(CurSyllable)，则实例化后可获取表 4 中全部特征值，如：Syllable+2 为“བཤད་”，Case+1 为“N”。

对以上的单一特征进行复合，得到以下复合特征模板，如表 5 所示。

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

表 5 复合特征模板

编号	模板
11	CurSyllable,Syllable-1
12	CurSyllable,Syllable+1
13	Syllable-1,Syllable+1
14	CurCas,Case-1
15	CurCase,Case+1
16	Case-1,Case+1

5 实验与分析

5.1 藏语句法标记识别结果

在句法标记识别中，由于以单音节为单位进行识别，因此正确率、召回率和 F 值均相同，因此本文仅使用正确率来表示黏写形式识别结果，如表 6 所示。

表 6 句法标记识别结果

	整体	S	R	V	N	C	M
P	0.98	0.95	0.85	0.93	1.00	0	0.93

黏写形式以及非黏写形式的句法标记的识别整体效果比较好。但是对于 R 类型的识别不是很理想，原因在于 ར(-r)后加字符构成音节字的出现频率较高，难以判断是音节固有的字符还是黏写字符，如：དར་、ཚར་、དཔར་、དྲར་等。其他的黏写形式也都可能存在与实词同型的情况，但从频率上看都不如 ར(-r)构成的黏写形式多。

5.2 基于音节的藏语功能组块边界识别结果

为了验证句法标记识别对基于 CRF 的藏语功能组块边界识别结果的影响，本文实现了两种条件下的基于 CRF 的边界识别实验，实验(1)为不进行文本预处理而直接以“·”对原始语料进行切分，进而对功能组块边界进行识别，作为 baseline；实验(2)对句法标记进行识别，并在此基础上以音节为单位对组块边界进行识别。实验结果如图 3 所示。

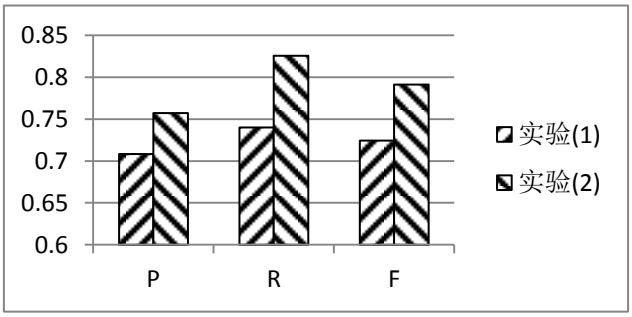

图 3 藏语功能组块边界识别实验

通过图 3 的实验结果对比，我们可以发现句法标记识别对于音节粒度上的功能组块边界识别效果有着明显的促进作用，在实验(2)中 F 值达到 79.12%，较 baseline 提高了 6.71%，说明了句法标记所隐含的语义信息确实对识别功能组块的边界有着促进作用，进一步验证了本文提出的预处理的有效性。

5.3 错误分析

直接在原始语料或者经初步预处理的原始语料基础上进行句法功能块的识别具有相当的难度，从我们的实验结果来看，尽管有一定的效果，但是还出现了不少的错误，这些错误归纳起来有如下一些类型：

(1)非谓语动词结构边界识别错误。非谓语动词结构是由短语或小句带有名词化标记后充当句子的某个句法功能块。它们是典型的长距离组块，识别难度较大。例如：

ཤོག་ལེ་རྩེ་རྒྱུ་ནི་སྤྲོ་སྣང་སྐྱེས་བའི་བྱ་བ་ཞིག་རེད།（shog phe rtse rgyu ni spro snang skyes bavi bya ba zhig red.玩纸牌是一件十分惬意的事。）识别结果为[ཤོག་ལེ་རྩེ་][རྒྱུ་][ནི་][སྤྲོ་སྣང་སྐྱེས་བའི་བྱ་བ་ཞིག་][རེད།]其中[ཤོག་ལེ་རྩེ་][རྒྱུ་]应为[ཤོག་ལེ་རྩེ་རྒྱུ་]，名词化 རྒྱུ་ 属于前一个块内。这类错误占据了整个错误的主要部分，是后续研究重点关注的对象。

(2)连谓结构边界识别错误。例如：ཁོ་མོ་མར་ལྷུང་སྟེར་མས།（kho mo mar lhung ster mas,她掉下来受伤了。）其中的 མར་ལྷུང་སྟེར་མས 应为一个块，但是被识别为[མར་ལྷུང་][སྟེར་][མས]三个块。

(3)同位结构或者缺乏标记的偏正结果边界识别错误，多个音节构成的同位结构或偏正结构由于缺乏显性标记，错误比较多。例如：ཁྱོད་ཚང་ལ་མི་དུ་ཡོད།(khyod tshang la mi du yod, 你家里有几口人？)其中 ཁྱོད་ཚང་ 缺乏定语标记，实验结果把它们处理为两个独立的块。

(4)缺乏时体态标记的光杆核心谓语块识别错误。藏语的动词（部分形容词）充当谓语居于句尾，在动词或形容词直接煞尾的句子中，动词之后缺乏足够的语言特征，从而使训练的模型的置信度不高，导致识别错误。例如：གཞུང་དྲང་པའི་མིས་གཏན་ནས་རྫུན་མི་བཤད།（gzhung drang pavi mis gtan nas rdzun mi bshad.品行好的人根本不说谎话。）其中 རྫུན་མི་བཤད་ 为谓语块识别错误。

尽管存在如上的一些错误，但是我们认为如果进一步细化识别特征，优化识别策略可以有效地改进识别结果。

6 结束语

功能组块代表了句子的各个功能性成分，使待分析句子的结构得以简化，在进行句法分析时大大降低了分析难度，解决了直接在分词、标注的基础上进行句法分析时由于词的数量较多产生的错误和歧义，从而导致分析结构的精确率低的问题。本文首先基于藏语句法功能组块的描述体系，提出了一种基于音节的藏语功能组块边界识别方法，通过实验分析以及结合藏语的语言特点，进一步将藏语句法标记识别作为预处理手段加入实验，最终结果准确率、召回率与 F 值分别达到 75.70%、82.54%与 79.12%。在下一步的工作中，我们一方面准备选取更优的特征以提高句法标记预处理的识别效果，另一方面，我们准备进一步扩大训练语料，在提升边界识别准确率的基础上，对功能组块的类型进行识别，为机器翻译等自然语言处理应用提供基础支持。

参考文献

John, D Lafferty, Andrew McCallum, Fernando CN Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data[A].2001.282-289

Kang, Caijun, Congjun Long, Di Jiang, Tibetan word segmentation based on word-position tagging, Asian Language Processing (IALP), 2013 International Conference on Digital Object Identifier: 10.1109 /IALP.2013.74 Publication Year: 2013 , Page(s): 239 – 242.

Long, Congjun, Caijun Kang, Di Jiang，The Comparative Research on the Segmentation Strategies of Tibetan bounded-variant forms，Asian Language Processing (IALP), 2013 International Conference on Digital Object Identifier:10.1109/IALP.2013.75 Publication Year: 2013 , Page(s): 243 - 246.

康才畯, 龙从军, 江荻.基于词位的藏文黏写形式的切分[J].计算机工程与应用 ISTIC PKU，2014，(11). DOI:10.3778/j.issn.1002-8331.1302-0075.

李琳, 龙从军, 江荻.藏语句法功能组块的边界识别[J].中文信息学报, 2013，(6):165-168.

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，(4):52-58

刘汇丹.2012.藏文分词及文本资源挖掘研究, 中国科学院大学, 博士论文。