# 基于错误驱动学习策略的藏语句法功能组块边界识别

王天航[1]，史树敏[1,2]，龙从军[3]，黄河燕[1,2]，李琳[3]

(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）

**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，为后续句子级深入分析提供支持。根据藏语的语言特点，本文在藏语句法功能组块描述体系基础上，提出基于错误驱动学习策略的藏语功能组块边界识别方法。具体思路为，首先基于条件随机场（Conditional Random Fields，CRFs）识别组块，然后分别基于转换规则的错误驱动学习（Transformation-based Error-driven Learning，TBL）及基于新特征模板的 CRFs 错误驱动学习进行二次识别，并行完成对初次结果的校正，F 值分别提高了 1.65%、 8.36%。最后通过实验分析，进一步将两种错误驱动学习机制融合，在 18073 词级的藏语语料上开展实验，识别性能进一步提高，准确率、召回率与 F 值分别达到 94.1%、94.76%与 94.43%，充分验证了本文提出方法的有效性。

**关键词：**错误驱动学习；藏语句法功能组块；组块边界识别；CRFs；TBL

**中图分类号：**TP391　　　　**文献标识码：**A

# Tibetan Functional Chunks Boundary Recognition Based on Error-Driven Learning Strategy

WANG Tianhang[1], SHI Shumin[1,2], LONG Congjun[3], HUANG Heyan[1,2],LI Lin[3]
(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)

Abstract: Tibetan syntactic functional chunk parsing is aimed at identifying syntactic constituent in Tibetan sentences that it can facilitate further analysis of sentences. According to the unique characteristics of Tibetan, the paper puts forward an error-driven learning strategy to identify the chunk boundary which is based on the description system of Tibetan syntactic functional chunk. The specific idea is as follows: we recognize the chunk boundary using the Conditional Random Fields (CRFs) model firstly. Then revise the recognition result through Transformation-based Error-driven Learning (TBL) method and the CRFs error-driven method. The F values of them increase 1.65% and 8.36%, respectively. Through the analysis of the experiments, we further combine these two error-driven techniques. In the experiment of the Tibetan corpus which contains 18073 words, the precision, recall and F value achieves 94.76%, 94.1% and 94.43% which proves that our method is effective.

**Key words:** error-driven learning; Tibetan syntactic functional chunk; chunk boundary recognition; CRFs; TBL

---

**收稿日期：**　　　　　　**定稿日期：**
**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

句法分析是自然语言处理领域的关键技术之一，在机器翻译、自动问答、信息抽取等诸多领域有着广泛应用。但是自然语言所具有的复杂性和不确定性，目前的完全句法分析器在性能与效率上还无法达到令人满意的效果。而采用“分而治之”的方法进行浅层的句法分析可以降低完全句法分析的难度。因此，组块分析逐渐成为目前的研究热点。

句法功能组块分析是组块分析的一个重要组成部分。它的研究目的是正确标注出构成句子的句法功能组块，自顶向下地进行句子拆分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，为进一步的事件骨架树分析提供最小的功能组块描述序列。随着藏语词法研究的深入，藏语语料库的建设、藏语分词工具的开发和藏语组块理论的提出都为进行藏语功能组块分析的研究奠定了基础。目前，已经初步具备了开展藏语功能组块分析研究的条件。

本文在项目组前期研究成果的基础上，首先采用 CRFs 模型利用改进的特征模板对组块边界进行初步识别；进而分别采用基于错误驱动的规则学习方法和基于错误驱动的统计学习方法对初次的识别结果进行校正，从而得出最终的识别结果。通过对实验结果错误分析，本文进一步提出了将基于错误驱动的规则与统计相结合的学习方法，进一步提高了对功能组块边界的识别效果。

# 2 相关工作

英语、汉语在组块上的研究成果较多[1-10]，这些成果为藏语功能组块的研究提供了较好的借鉴。

国内藏语组块理论及识别也已经开展。依据藏语的句法形式标记、词类以及其句法分布特点，文献[11]定义了 7 类句法组块。文献[12]对 7 类分类体系进行了扩充，提出了 10 类藏语句法组块。文献[13]对谓语组块的组成成分进行了深入研究，确定了 9 类体貌-示证标记范畴及形式标记。文献[14]针对现代藏语名词短语开展了研究，根据名词组块的三类句法形式标记构造识别规则集预测名词组块边界。文献[15]提出了识别部分谓语组块的语法规则和算法，详细分析了藏语助动词的语法功能。这些研究的主要技术路线都是基于藏语句法规则进行的，目前采用统计方法对藏语组块进行识别的研究还较少。而随着藏语语言信息化程度的提高，近期也有学者从工程应用角度尝试用统计的方法对藏语组块进行识别。文献[16]是从汉语组块出发，通过对译译文寻找藏语组块的边界。但是该研究没有充分考虑藏语自身特点，特别是格标记属性特征的作用。文献[17]界定了 5 种句法功能组块，并首次将 CRFs 模型运用到对藏语功能组块边界的识别，但选取的特征模板相对简单。

# 3 藏语功能组块

## 3．1 藏语功能组块描述性定义

本文所使用的功能组块定义遵循前期研究中采用的描述性定义，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[17]。

## 3．2 藏语功能组块标注集

本文采用 *BIE* 标记集来标记功能组块，使得句法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“ངའི་བོད་ཡིག་དགེ་རྒན་བློ་བཟང་ལགས་རེད།”（我的藏文老师是洛桑拉，ngavi bod yig dge rgan blo bzang lags red）为例，利用该标记集对其进行标记的中间结果为：

[ང/rhའི་/wgབོད་ཡིག་/ngདགེ་རྒན་/ng][བློ་བཟང་/nhལགས་/z][རེད/vl][།/xp]进一步处理后得到的标注结果见图 1。

Eg1: ང/B འི་/I བོད་ཡིག་/I དགེ་རྒན་/E བློ་བཟང་/B ལགས་/E རེད/B །/B

Latin: nga/B vi/I bod yig/I dge rgan/E blo bzang/B lags/E red/B ./B

例 1：我的藏文老师是洛桑拉。

图 1 藏语功能组块边界识别标注实例

# 4 边界识别统计模型

## 4．1 前期工作

在前期工作中我们采用前后各两个词及当前词的词形和词性以及前一个词和当前词的转移概率特征作为特征利用 CRFs 尝

试了藏语功能组块边界识别，实验结果 *F* 值达到 83.56%。为了进一步确定更优的特征，本文通过基于信息增益的特征选择实验，发现前后词的词形以及当前词的词音节数所蕴含的信息对藏语功能组块的边界识别有着促进的作用，并且通过进一步的实验，将以上的单一特征进行组合，提出了改进的边界识别特征模板。

## 4．2 改进的 CRFs 特征模板

好的模板有助于对功能组块边界识别。通常来讲，丰富的上下文特征对于边界的识别率提高有着积极的作用，但过量的特征反而可能会降低训练的效果，并且会使训练和测试过程开销大大增加。考虑到这些因素，本文定义了如下原子模板，如表 1 所示。

表 1 原子特征模板

| 编号 | 模板 | 编号 | 模板 |
|---|---|---|---|
| 1 | *CurPOSTag* | 8 | *POSTag+2* |
| 2 | *CurWord* | 9 | *Word-1* |
| 3 | *ChunkTag-1* | 10 | *Word-2* |
| 4 | *ChunkTag-2* | 11 | *Word+1* |
| 5 | *POSTag-1* | 12 | *Word+2* |
| 6 | *POSTag-2* | 13 | *CurRhythm* |
| 7 | *POSTag+1* | | |

其中，*POSTag* 表示词性标注，*ChunkTag* 表示组块标记，*Word* 表示词形，*Rhythm* 表示词音节数，实验窗口取 5，上表中+/-表示当前词后/前的词对应的特征。当特征函数取特定值时，则该模板被实例化，如图 2 所示。

Eg2:[ང/rh][ས་/wa][གསར་འགྱུར་/ngབཤད་/vnམཁན་/hདེ་/rd][ངོ་ཤེས་/voཀྱི་ཡོད/t][།/xp]

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

例 2：我认识那个播音员。

图 2 原子模板特征选择示例

其中，以“གསར་འགྱུར་”（新闻，gsar vgyur），为当前词（*CurWord*），则实例化后可获取表 1 中全部特征值，如：*CurPOSTag* 为“ng”，*Word+1* 为 “བཤད་”。

对以上的单一特征进行复合，通过实验选择，得到以下对边界识别有益的复合特征模板，如表 2 所示。

表 2 复合特征模板

| 编号 | 模板 |
|---|---|
| 14 | *ChunkTag-1,ChunkTag-2* |
| 15 | *ChunkTag-1,ChunkTag-2,CurPOStag* |
| 16 | *ChunkTag-1,ChunkTag-2,CurWord* |
| 17 | *POSTag-1,POStag-2* |
| 18 | *POSTag-1,POStag-2,CurPOS* |
| 19 | *POSTag-1,CurPOS,POSTag+1* |

## 5 基于错误驱动的组块边界识别策略

错误驱动是一种通过对错误标记的上下文特征进行学习从而将错误标记校正为正确标记的学习算法，这种算法在词性标注和句法分析中都取得了不错的效果。本文分别利用规则的错误驱动技术与统计的错误驱动技术对功能组块边界进行识别校正

## 5．1 基于转换规则的错误驱动学习

TBL 算法是 Brill 于 1992 年提出的一个有效的学习算法 [19]，它的核心任务是构建用于校正的转换规则集。利用 TBL 自动获取转换规则集的算法如下：

**算法 1**：基于转换规则的错误驱动学习算法

*Input*: TC (训练语料) RT（规则转换模板集合）

*Initialize*: RS=∅，Score=∅，EF，TM（RS 为规则集合（有序），Score 为对应的规则评价得分，EF 为评价函数，TM 为初始标注器）

*Step1*:利用 TM 对 TC 进行初始标注，得到初始标注结果 Result。

*Step2*:从 RT 中取出一条未选规则模板 $rt_i$ 带入 Result，通过错误学习得到转换规则集合 TRS，并将 $rt_i$ 标记为已选。

*Step3*:从 TRS 中取出一条规则 $r_i$ 对 TC 中满足规则触发条件的标记进行转换，利用评价函数对 $r_i$ 对其打分得到评分 $s_i$，将 $s_i$ 加入 Score。

*Step4*:重复 *Step2,Step3* 直到 RT 中所有规则模板均已选，取出对应评分最高的 $r_i$ 加入到 RS，利用 $r_i$ 对 TC 进行标注得到新的标注结果 Result，将 RT 中所有规则模板重新初始化, Score 清空。

*Step5*:重复 *Step2,Step3* 直到所有 $r_i$ 对应评分均为负。

*Output*: RS

转换规则模板可以利用的属性信息包括：词形信息（*Word*），词性信息（*POS*），功能组块标注信息（*ChunkTag*）。通过以上三类信息的组合使用，本文构造了以下的模

板转换条件，如表 3 所示。

表 3 TBL 规则转换模板

| 编号 | 模板转换条件 |
|---|---|
| 1 | *ChunkTag-1,CurWord* |
| 2 | *POSTag-1,CurPOSTag* |
| 3 | *POSTag-1,POSTag+1* |
| 4 | *POSTag-2,POSTag-1,CurPOSTag* |
| 5 | *ChunkTag-2,ChunkTag-1,CurWord* |
| 6 | *POSTag-1,CurPOSTag,POSTag+1* |
| 7 | *CurPOSTag,POSTag+1,ChunkTag-1* |
| 8 | *ChunkTag-2,ChunkTag-1,CurPOSTag* |
| 9 | *ChunkTag-1,POSTag-1,CurPOSTag* |

实验采用的评价函数如下：

$$F = X_R(r) - X_E(r) \qquad （1）$$

其中，$F$ 为转换规则的评价函数，$X_R(r)$ 为应用规则 $r$ 后正确的标记数，$X_E(r)$ 为应用规则 $r$ 后错误的标记数。通过上述过程，就可以得到用来校正的转换规则集。

### 5．2 基于 CRFs 的错误驱动学习

与传统的规则方法和基于转换的错误驱动方法不同，基于 CRFs 的错误驱动技术是由统计模型来自动驱动纠错。它是将第一阶段 CRFs 本身识别的结果作为一般特征加入下一阶段的 CRFs 特征模板中，并融入第一阶段所应用的特征进行二次识别。CRFs 错误驱动学习可以利用的属性信息包括当前词第一阶段的组块标记结果（*Tag*）与原模板特征集（*POS, Word, ChunkTag*）。根据以上属性信息的组合使用，本文构造了以下特征模板，如表 4 所示。

表 4 CRFs 错误驱动学习特征模板

| 编号 | 模板 |
|---|---|
| 1 | *Tag, ChunkTag-1,ChunkTag-2* |
| 2 | *Tag, CurWord, CurPOSTag* |
| 3 | *Tag, Tag-1, Tag-2* |
| 4 | *Tag, Tag-1, Tag+1* |
| 5 | *Tag, POSTag-1,POSTag+1,CurPOS* |

以“སྦྲང་མ་/ng མེ་ཏོག་/ng གི་/wgགསེབ་/nl ནས་/wc འཕུར་ལྡིང་བྱས/vo །/xp”（蜜蜂在花丛中飞舞，Spramg ma me tog gi gseb nas vphur lding byas）为例，通过初次的 CRFs 识别得到标注结果[སྦྲང་མ་/མེ་ཏོག་/གི་/གསེབ་][ནས་] [འཕུར་ལྡིང་བྱས][།]，但正确结果应该为 [སྦྲང་མ་][མེ་ཏོག་/གི་/གསེབ་] [ནས་] [འཕུར་ལྡིང་བྱས][།]。因此需要对这种错误进行校正。以“གི་”为当前词，表 5 给出了基于 CRFs 的错误驱动技术的特征选择。

表 5 基于 CRFs 错误驱动技术的特征选择

| 位置 | 词 | POS | CRFs 标注 | 正确标注 |
|---|---|---|---|---|
| -2 | སྦྲང་མ་ | ng | B | B |
| -1 | མེ་ཏོག་ | ng | I | B |
| 0 | གི་ | wg | I | I |
| 1 | གསེབ་ | nl | E | E |
| 2 | ནས་ | wc | B | B |

将以上的特征带入表 4 对应的模板，就可以得到用来进行 CRFs 错误驱动的模板实例。

## 6 实验

### 6．1 实验数据及评价参数

实验使用 Taku Kudo 开发的开源 CRFs++[1]进行模型训练。选择 fnTBL[2]作为转换规则错误驱动工具。实验语料共包含 18073 个词，将其随机分为三部分，第一部（7498 个词）用来进行初次的 CRFs 训练，第二部分（8043 个词）用来错误驱动学习，第三部（2532 个词）作为最终的测试数据，用来测试训练效果。实验采用以下评价标准：

$$P = \frac{正确功能组块数}{召回组块总数} \times 100\% \qquad （2）$$

$$R = \frac{正确功能组块数}{功能组块总数} \times 100\% \qquad （3）$$

$$F = \frac{2 \times P \times R}{P + R} \times 100\% \qquad （4）$$

### 6．2 实验过程与结果分析

#### 6．2．1 错误驱动学习方法独立实验

对训练数据进行预处理后运用 4.2 所提出的特征模板对训练语料进行训练，得到初步训练模型，分别利用两种错误驱动学习方法对初步识别结果进行二次训练，得到最终的模型，并在测试数据中运用上述两种训练模型进行测试得出最终的功能组块识别结果，如图 3 所示。

---

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

[2] http://nlp.cs.jhu.edu/~rflorian/fntbl/index.html

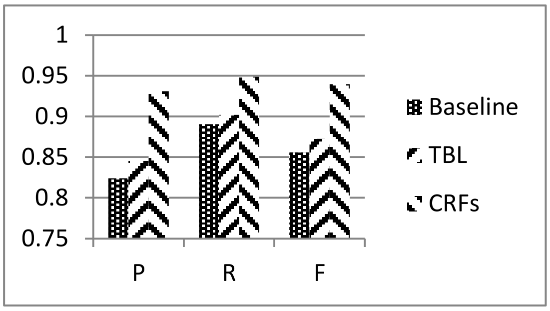

图 3 规则与统计错误驱动学习方法独立实验结果

图中Baseline表示CRFs初次识别结果，TBL 表示在初次识别后运用 TBL 错误驱动后的识别结果，准确率、召回率与 *F* 值分别达到 84.54%、90.18%与 87.27%，较初次识别 *F* 值提高了 1.65%。CRFs 表示在初次识别后运用 CRFs 进行错误驱动的识别结果，准确率、召回率与 *F* 值分别达到 93.11%、94.86%与 93.98%，较初次识别 *F* 值提高了 8.36%。

通过 CRFs 错误驱动的校正，已经获得了不错的边界识别效果，但是在对识别错误分析后，发现对长距离功能组块的边界识别出错率较高，CRFs 模型会将一些距离较长的功能组块切割成几个小的功能组块，而正确的功能组块开始边界与结束边界可以与 CRFs 识别后的功能组块的部分边界对应，以“སློབ་ཞུགས་/voགོང་/ng གི་/wg སློབ་གསོ་/ng དང་ལེན་བྱ་/vo རྒྱུ/h འི་/wg འགན་ལེན་/ng བྱས་/vo ཡོད་/t པ་རེད/t །/xp”为例，正确的组块边界识别结果应为 སློབ་ཞུགས་/B གོང་/I གི་/I སློབ་གསོ་/I **དང་ལེན་བྱ་/I** རྒྱུ/I འི་/I འགན་ལེན་/E བྱས་/B ཡོད་I པ་རེད/E །/B。经过 CRFs 错误驱动校正后的边识别结果为 སློབ་ཞུགས་/B གོང་/I གི་/I སློབ་གསོ་/I **དང་ལེན་བྱ་/B** རྒྱུ/I འི་/I འགན་ལེན་/E བྱས་/B ཡོད་I པ་རེད/E །/B。因为块中“དང་ལེན་བྱ་”的错误识别，将正确的功能组块切分成为两个错误的小功能组块，因此只要想办法把功能组块中“དང་ལེན་བྱ་”所对应的错误识别标记 *B* 校正为 *I* 即可得到正确的功能组块边界识别结果

### 6．2．2 错误驱动学习方法融合实验

通过对上面 CRFs 错误驱动的实验结果分析，本文进一步提出设想，能否运用规则的方法进一步对识别结果校正，从而弥补 CRFs 模型对长距离功能组块边界识别效果较低的缺陷。于是将第二部分训练语料分为两部分 $T_1$，$T_2$，在都经过 CRFs 初步训练之后，将 $T_1$ 用来进行 CRFs 错误驱动学习，并将训练结果应用于 $T_2$ 之上，最终再在 $T_2$ 上利用 TBL 对其进一步训练，从而将统计与规则错误驱动学习机制相结合，识别结果如图 4 所示。

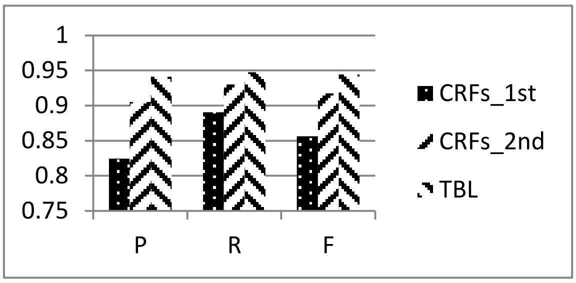

图 4 规则与统计错误驱动学习方法融合实验结果

图中 CRFs_1st 表示 CRFs 初次识别结果，CRFs_2nd 表示经过初步识别后利用 CRFs 进行错误驱动学习的识别结果，TBL 表示在前两步的基础上利用 TBL 进行错误驱动学习的最终结果，准确率、召回率与 *F* 值分别达到 94.1%、94.76%与 94.43%，与前次实验结果相比，*F* 值提升了 0.45%，实验结果表明通过两种错误驱动学习方法的结合能够提升对功能组块边界特别是长距离功能组块边界的识别效果，尤其是在实验所用训练语料中长距离的句法功能组块不多且语料的总体量有限的情况下。

### 6．3 实验结果分析

通过上述实验可见，利用本文所提出的方法能够有效地对功能组块边界进行识别，但仍然存在一些典型的边界识别错误，本文将其归纳如下：

(1)常用副词位置不固定导致识别错误：藏语中一些副词位置不固定，既可以放在修饰词前也可以放在修饰词后，遇到这种情况，往往会出现错误。

例 3:[ཇོ་མོ་གླང་མ་/nz][ནི་/up][རི་བོ་/ngཧ་ཅང་/dཆེན་པོ་/aཞིག་/m][རེད/vl][།/xp]

珠穆朗玛是非常高的山。

例 3 中由于 ཧ་ཅང་(非常)与 ཆེན་པོ་(大)的搭配既可以为ཧ་ཅང་ཆེན་པོ་ 也可以为ཆེན་པོ་ཧ་ཅང་ 其先后顺序多变，在本文进行边界识别时，往往把 ཧ་ཅང་ 错误的识别为一个新的功能组块的开始。

(2)名词化标记导致识别错误：藏语中非谓动词后常常紧跟名词化标记，使整个非谓

动词短语充当某种成分，遇到这种情况，往往会出现识别错误。

例 4:[ཁོ་མོ་/rh][ནངས་སྔ་མོ་/nt][ལངས་/voརྒྱུ་/h][ར་/wd][མི་/dnདགའ/vo][།/xp]

她不喜欢早起。

例 4 中名词化标记 རྒྱུ་ 紧跟动词 ལངས་(起)之后，使名词化短语充当 དགའ 的对象宾语，应与 ལངས་ 划分为同一块内，但被错误的分成两个独立块。

(3)疑问语气词处于句中导致识别错误：通常情况下，疑问语气词会出现在句子的末尾，但有时由于出现嵌套句，疑问语气词出现在嵌套小句的末尾从而被嵌入句中，遇到这种情况，往往会出现识别错误。

例 5:[ང/rh][ས་/wa][ཞོགས་པ/nt][ར་/wl][བདེ་མོ་/a][ཡིན་/vl][ནམ་/y][ཞེས་/qt][དྲིས་/vo][ཀྱང་/c][ཁོ་མོ/rh][ས་/wa][ཕྱིར་/dའཚམས་འདྲི་/ng][མ་/dn][བྱས/vo][།/xp]

我问：“早上好吗?”，但她不回答。

例 5 中，疑问词 ནམ་ 由于是嵌套小句的末尾而被嵌入句子中，应为小句谓语组块，但被错误地划分在主句谓语组块内。

## 7 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的难度，也大大避免了直接在分词的基础上进行句法分析时由于词的数量较多引起的歧义，从而导致分析结构的精确率低等缺点。本文首先基于前期工作中藏语句法功能组块的描述体系，提出了一种错误驱动学习策略与条件随机场相结合的藏语功能组块边界识别方法，通过实验分析，进一步将两种错误驱动学习方法融合，最终实验结果准确率、召回率与 *F* 值分别达到 94.1%、94.76%与 94.43%。在下一步的工作中，我们准备在边界识别的基础上，进一步对功能组块的类型进行识别。

## 参考文献

[1] Abney, Steven P.“Parsing by chunks”. Springer Netherlands, 1992.

[2] Ramshaw, Lance and Mitchell Marcus. Text Chunking using Transformation-Based Learning. In Proceedings of the ACL Third Workshop on Very Large Corpora, June 1995. 82-94.

[3] Tjong Kim Sang E F, Buchholz S. Introduction to the CoNLL-2000 shared task: Chunking[C]//Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning-Volume 7. Association for Computational Linguistics, 2000: 127-132.

[4] Pierce D, Cardie C. Limitations of co-training for natural language learning from large datasets. The 2001 Conference on Empirical Methods in Natural Language Processing, Cornel University, Ithaca NY, 2001:1-9.

[5] 李衎, 朱靖波, 姚天顺. 基于 SVM 的中文语块分析[J]. 中文信息学报, 2004, 18(2), 1-7.

[6] 李素建, 刘群, 杨志峰. 基于最大熵模型的组块分析[J]. 计算机报, 2003:1722-1727.

[7] Tan Y M, Yao T S, Chen Q et al. Applying conditional random fields to Chinese shallow parsing. Proceedings of CICLing2-2005. Mexico City, Mexico, 2005: 167-176.

[8] 周强, 赵颖泽. 汉语功能块自动分析[J]. 中文信息学报,2007,21(5):18-24.

[9] 陈亿, 周强, 宇航. 分层次的汉语功能块描述库构建分析[J]. 中文信息学报,2008,22(3): 24-31.

[10] 黄德根，于静. 分布式策略与 CRFs 相结合识别汉语组块[J]. 中文信息学报，2009，23(1)：16−22.

[11] 江荻. 现代藏语的句法组块与形式标记[A]. 语言计算与基于内容的文本处理, 孙茂松, 陈群秀主编. 北京:清华大学出版社. 2003:160-166.

[12] 江荻. 面向机器处理的现代藏语句法规则和词类、组块标注集. 江荻、孔江平主编, 中国民族语言工程研究新进展, 北京：社会科学文献出版社， 2005, 13-106.

[13] 江荻. 藏语拉萨话的体貌、示证及自我中心范畴[J]. 语言科学. 2005, (1):70-88.

[14] 黄行, 孙宏开, 江荻, 张济川, 唐黎明. 现代藏语名词组块的类型及形式标记特征[A]. 孙茂松, 陈群秀(主编):自然语言理解与人规模内容计算[C]. 清华大学出版社. 2005, 615-618.

[15] 龙从军, 江荻. 现代藏语带助动词谓语组块的识别方法[A]. 第 2 届青年计算语言学会议论文[C]. 2004.

[16] 诺明花， 刘汇丹， 马龙龙， 等. 基于中心语块扩展的汉藏基本名词短语对的识别[J]. 中文信息学报， 2013， 27(4)： 63-69.

[17] 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， 27(6).

[18] Lafferty, John, Andrew McCallum, and Fernando CN Pereira. “Conditional random fields: Probabilistic models for segmenting and labeling sequence data.” (2001)

[19] Brill E. Transformation-based error-driven parsing[C]//Proceedings of the Third International Workshop on Parsing Technologies, Tilburg, The Netherlands. 1993.