藏语句法功能组块的边界识别

李琳
中国社会科学院研究生院
中国社会科学院民族学与人类学研究所
lilin20081@foxmail.com

龙从军
中国社会科学院民族学与人类学研究所
中央民族大学民族语言监测分中心
longcj@cass.org.cn

江荻
中国社会科学院民族学与人类学研究所
jiangdi@cass.org.cn

摘要：藏语句法功能组块能够很好地描述藏语句子的基本骨架，是连接句法结构与语义描述的重要桥梁。根据藏语句法特点，本文作者提出五种句法功能组块及功能组块边界识别策略。文章首先描述了藏语句法功能组块的基本特点和标注体系，然后在此基础上提出了一种基于条件随机域(CRFs)模型的功能组块边界识别算法。小规模训练语料的实验结果表明，该方法可以有效的识别出功能组块边界，值得进一步研究。

关键词：藏语句法功能组块；组块边界识别；条件随机域模型

中图分类号：TP391　　**文献标识码：**A

Tibetan Functional Chunks Boundary Detection

Abstract: Tibetan functional chunks describe a sentence skeleton, and they are the link between sentence structure and semantics. In this paper, we proposed primary functional chunks in Tibetan and a functional chunk tag system. Based on the theory, a functional chunk boundary detection algorithm was proposed. Experiments on a limited scale data suggest that the algorithm is capable of recognizing most boundaries and deserves to be studied deeply.

Key words: Tibetan functional chunks; chunks boundary detection; CRFs

1 引言

句法分析是自然语言处理的基础技术，被广泛地应用到机器翻译、信息抽取等诸多研究领域。目前句法分析技术的一个重要发展趋势是由完全句法分析转向部分句法分析的研究。基于块的部分句法分析可以降低句法分析的复杂性，提高局部分析的准确性，从而为进一步的完全句法分析和语义分析奠定基础。英汉句法分析的研究成果很多[1-10]，尤其是组块边界识别的研究为藏语组块边界的识别提供了较好的经验和技术积累。

对藏语句法组块理论及识别方法探讨已经有了较多成果。文献[11]从藏语的高层单位(短语、结构及句法成分组)切入，提出了现代藏语句法特征的组块描述体系，该系统包括八种类型的藏语句法组块。在此基础上，文献[12]对该描述体系进行了扩充。文献[13]通过藏语助动词的句法分布特征，探讨了识别带助动词的谓语组块；文献[14]讨论了藏语形容词谓语句的谓语结构和形式标记并进行实验；文献[15]提出了依靠右边界的名词组块识别方法；文献[16]提出了识别藏语判定句的主语和宾语的方法;文献[17]讨论了藏语述说动词句宾语的识别方法。

目前对藏语组块的研究主要是采用规则的方法对不同类型的组块进行识别。在前人对藏语句法组块的描述基础上，本文提出了基于 CRFs 模型的藏语功能组块边界识别方法。从实践角度对藏语功能组块边界进行识别，并对错误结果进行分析，为进一步的组块边界识别与组块类型标注积累经验。

2 藏语功能组块体系

功能组块描述体系是自顶向下描述句子的基本骨架[10]，在该体系中描述单元可以是句子层面的谓词和与它相关联的体词，如谓词与各种论元。由于藏语具有丰富的句法标记，描

述单元之间的关系更加清晰，因此能够借鉴英汉组块识别的方法，从高层语言单位切入分析藏语句法结构和句法功能组块。

现代藏语总的语序是主语-宾语-谓语，表达完整意义的扩展句法语序是：主语+（间接宾语）+（直接宾语）+（结果补语）+（状语）+动词+（状态补语）[12]。从句法成分的各个位置上看，藏语句子中与句法组块存在对应关系的句法成分有主、宾、谓、状、补[1]，名词或体词的修饰语组块未单独列出。根据这些研究成果，本文建立了藏语功能组块描述体系，如表1所示。

表 1. 藏语功能组块描述体系

功能块标记	功能块描述
S	主语块
P	谓语块
O	宾语块
D	状语块
C	补语块
M	句法标记

3 基于 CRFs 的藏语功能组块识别

3．1 藏语功能组块标注集

为了将识别功能组块边界问题转化为序列标注问题，本文采用 Start/End 标记集[18]来标记功能组块。标记集中的每个标记均由两部分构成：第一部分是词语所属功能组块的类型标记，具体如表 1 所示；第二部分为该词语在功能组块中的位置，起始位置用 B 表示，内部位置用 I 表示，结束位置用 E 表示，只包含一个词的块用 U 表示；在这两部分标记之间用“-”来分隔。对于不属于这几类功能组块的单词和符号，统一使用 N 来标记。

以“ང/rhས/waཁ་ལག/ngབཟོ/voཀི་ཡོད/t།/xp”（我在做饭。）为例，利用该标记集对其进行标记的中间结果为：[S ང/rh][M ས/wa][O ཁ་ལག/ng][P བཟོ/voཀི་ཡོད/t]།/xp。再经过处理后得到的标注结果见表 2。

表 2. 藏语功能块标注实例

词语	词性	标记
ང	rh	S-U
ས	wa	M-U
ཁ་ལག	ng	O-U
བཟོ	vo	P-B
ཀི་ཡོད	t	P-E
།	xp	N

[1]定语与中心语之间的标记不能或者极少作为组块边界标记，因此本文不单独列示。

3．2 条件随机域模型

藏语功能组块边界识别问题可以转化为序列标注问题，本文利用 CRFs 模型建立功能组块的序列标注模型。CRFs 模型是一个基于无向图的条件概率模型，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。CRFs 模型在各类序列标注问题中都显示出了很好的处理效果，如词类标注、专有名词识别、语义角色标注等。选择 CRFs 模型是因为它能够任意添加有效的特征向量，从而综合利用词、词性等多层次的语言信息。

具体计算问题可以描述如下：设输入的序列为 $X=x_1x_2x_3\cdots x_n$，其中 x_i 为一个词语，并带有相应的词性标记，相应的输出序列为 $Y=y_1y_2y_3\cdots y_n$，其中 $y_i\in T$。则对一个输入序列 X 进行标注的过程就是为其寻找一个最优的输出标记序列 Y 的过程。

4 实验

4．1 实验语料及评价参数

实验使用 Taku Kudo 开发的开源 CRF++软件包 0.53。实验语料采用拉萨藏语口语语料库，该语料库标注了词性和功能组块的边界信息。由于标注语料较少，我们采用交叉验证的方式，将语料平均分为 4 份，进行了 4 次试验。试验结果是这 4 次试验数据的平均值。每次实验对语料按 8:2 进行划分，其中训练集包含 800 个句子，测试集包含 200 个句子。使用自然语言处理常用的评价方法对功能组块边界识别性能进行评价：

(1)准确率(Precision)：

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

(2)召回率(Recall)：

$$R=(\text{正确功能组块数}/\text{功能组块总数})\times 100\%$$

(3) F-1 测度(F-1 measure)：

$$F=(2\times P\times R)/(P+R)$$

4．2 特征模板

CRFs 模型识别功能块边界的关键在于特征的选择，其恰当与否会对识别结果产生直接的影响。通常来讲，丰富的上下文特征对于识别精确率的提高有着积极的作用，但会给训练和测试过程带来很大的开销。因此，应在保证实验效果的情况下，所选取的特征应尽可能少。本文在进行特征选择的时候，考虑到词和词性及其上下文之间存在着的种种依赖关系，尝试将当前位置的前后两个词及词性作为特征。这种组合包括了词和词性标记的组合信息，可以对模型提供更丰富的识别信息。本实验利用不同模板进行了分组实验，详见表 3。

表 3. 功能块边界识别特征模板

	特征说明
template1	前后各两个词及当前词
template2	前后各两个词及当前词和词性
template3	前后各两个词及当前词和词性 前一个词和当前词的转移概率特征
template4	前后各两个词及当前词和词性 后一个词和当前词的转移概率特征
template5	前后各两个词及当前词和词性 前一个词和当前词的转移概率特征 后一个词和当前词的转移概率特征

4．3 实验结果

利用表 2 的特征模板，利用训练语料对 CRFs 模型进行训练，再利用得到的模型对测试语料进行标注，最后得到功能组块边界识别结果。表 4 为在不同特征模板下训练的 CRF 模

型自动识别功能组块的效果。

表 4. CRFs 识别结果

	P(%)	R(%)	F(%)
template1	76.75	70.92	73.72
template2	78.55	75.37	76.93
template3	88.26	79.31	83.56
template4	81.07	77.13	79.05
template5	85.84	79.33	82.46

实验结果表明，采用 template3 的时识别模型效果最好，F 值达到了 83.56%。这比 template1 提升了 9.8%，比 template2 提升了 6.6%，说明前一个词和当前词转移概率特征的加入，使得系统能够识别出更多的功能块，尤其对功能块准确率的提高更为明显。template4 的实验效果不如 template3 好，这证明采用“后一个词和当前词的转移概率特征”比“前一个词和当前词的转移概率特征”效果好。虽然丰富的上下文特征能够提高模型的性能，然而 template3 的效果却比 template5 要好，这说明在某些情况下，过多的上下文特征，反而会使识别效果下降。

4．4 错误分析

在使用 CRFs 模型对功能组块边界进行识别后，错误率仍然较高，主要的原因有以下几个方面：

(1)复杂名词组块分析错误：藏语名词组块功能多样、结构复杂，尤其是遇到名词组块嵌套的情况，其识别结果往往出现错误。

例 1：[རེད་སང་/nt][བོད་/ng][ལ་/wx][ཡོང་/vo$_{མཁན་}$/h$_{ཕྱི་རྒྱལ་}$/ng][མང་པོ་/a][འདུག་/ve$_{གས་}$/y]།/xp

现在来西藏的人中有很多外国的。

例 2：[ཁྱ་ཀ་/ng$_{འདི་}$/rd][བཙའ་མཁན་/ng][ར་/rh$_{དེ་}$/wg$_{ཨ་མ་}$/ng$_{ལགས་}$/z][རེད་/vl]།/xp

做这个鸡肉的是我妈妈。

例 1 中，主语块བོད་ལ་ཡོང་མཁན་ཕྱི་རྒྱལ་是由名词化短语做定语修饰另一个名词构成的嵌套型名词短语。CRFs 模型的标记结果倾向于将较长的块切分为较短的块，因此错误地将བོད་误识别为一个单独的块。例 2 也能说明这种情况的普遍性，主语块ཁྱ་ཀ་འདི་བཙའ་མཁན་也是一个嵌套型的名词组块，བཙའ་མཁན་被错误地排除在前一个组块之外。

(2)比较句的识别错误：比较句的比较主体和比较客体可能都是名词性成分，如代词或名词，而比较值往往是形容词。因此在块边界识别时倾向于将比较客体和比较值点合在一起构成一个组块。这种类型的错误比较难处理。例如：

例 3：[ཨཇོན་ཁང་/ng$_{འདི་}$/rd][ལས་/wb][ཕ་གི་/rd$_{ཡག་གོ་}$/a][འདུག/va]

这个旅馆比那个好。

例 3 中ཕ་གི་作为比较客体应是独立的块，但被错误地与ཡག་གོ་划分在一个块内。

(3)由于可用的训练语料过少，数据稀疏问题影响了 CRFs 模型的识别效果。而且对于句

子结构的不同理解，也给标注造成一些不一致的情况。由于功能组块标注是采用人工标注，在工作中难免存在主观因素的影响。如果能够采用机器初步标注，后期再进行人工校对的方式，就可以避免标注手法不一致对结果的影响。

5 结束语

在以往研究的基础上，本文将 CRFs 模型引入藏语功能组块边界识别工作，尝试使用不同语言信息构造特征模板，进而构建不同的识别模型。实验结果表明，基于统计的方法在块边界识别中效果比较明显。在下一步工作当中，我们一方面要进一步扩大训练语料和确定更优的特征，另一方面可以引入错误驱动的方法对处理结果加以校正。

参考文献

[1] Abney S P. Parsing by chunks[M]. Springer Netherlands, 1992.

[2] 周俏丽, 刘新, 郎文静, 蔡东风. 基于分治策略的组块分析[J]. 中文信息学报, 2012, 26(5): 120-128.

[3] 周强. 汉语基本块描述体系[J]. 中文信息学报, 2007, 21(3): 21-27.

[4] 周俊生, 戴新宇, 陈家俊等. 基于大间隔方法的汉语组块分析[J].软件学报,2009, 20(4):870-877.

[5] 黄德根, 王莹莹. 基于 SVM 的组块识别及其错误驱动学习方法[J]. 中文信息学报, 2006, 20(6): 17-24.

[6] 周强,李玉梅.汉语块分析测评任务设计[J]. 中文信息学报, 2010, 24(1):123-128.

[7] 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.

[8] 李国臣, 王瑞波, 李济洪. 基于条件随机场模型的汉语功能块自动标注[J]. 计算机研究与发展, 2010 (002): 336-343.

[9] 刘海霞, 黄德根. 语义信息与 CRF 结合的汉语功能块自动识别[J]. 中文信息学报, 2011, 25(5): 53-59.

[10] 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.

[11] 江荻. 现代藏语组块分词的方法与过程[J]. 民族语文, 2003, 4: 31-39.

[12] 江荻. 面向及其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] 龙从军, 江荻. 现代藏语带助动词的谓语组块及其识别. 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, 123-135.

[14] Jiang Di, Hu Hong-yan. The construction and identification approaches of adjectival predicate in modern Tibetan[J]. Studies in Language and Linguistics, 2005, 25(2): 115-122.

[15] 黄行, 孙宏开, 江荻, 等. 现代藏语名词组块的类型及形式标记特征[C]//全国第八届计算语言学联合学术会议 (JSCL-2005) 论文集. 2005.

[16] 黄行, 江荻. 现代藏语判定动词句主宾语的自动识别方法[J]. , 孙茂松语言计算与基于内容的文本处理, 2003, 172.

[17] 江荻. 藏语述说动词小句宾语及其标记[J]. 中文信息学报, 2007, 21(4): 111-115.

[18] Manabu Sassano and Takehito Utsuro. Named Entity Chunking Techniques in Supervised Learning for Japanese Named Entity Recognition[C]. In Proceedings of COLING 2000: 705－711.