文章编号：1003-0077（2018）00-0000-00

# 基于句法树的藏语最长名词短语识别

龙从军[1] 刘汇丹[2] 周毛克[3]
（1.中国社会科学院民族学与人类学研究所，北京 100081；
2. 中国科学院软件研究所，北京 100190；
3.中国社会科学院研究生院，北京 100081）

**摘要：**最长名词短语携带着丰富的句法和语义信息，经常与句法成分对应，在句子中充当一定的语义角色。最长名词短语识别在自然语言处理中占重要地位，是分析和理解句子结构和意义的基础。本文通过梳理不同概念的最长名词短语的含义，从句法树角度界定了藏语最长名词短语的基本概念；从句法树库中抽取 6038 个句子，分析了最长名词短语的结构类型、边界特征和出现频次，最后采用序列标注模型和句法分析模型对最长名词短语进行识别。序列标注模型识别结果的正确率、召回率和 F1 值分别为 87.14%、84.72%、85.92%。句法分析模型识别结果的正确率、召回率、F1 值分别为 85.02%、84.51%、84.76%。

**关键词：**藏语句法树；最长名词短语；名词短语类型

**中图分类号：**TP391 **文献标识码：**A

# Recognition of Tibetan the longest noun phrases based on syntax tree

LONG Congjun[1], LIU Huidan[2], ZHOU Maoke[3]

(1. Institute of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing 100081, China
2. Institute of Software，Chinese Academy of Sciences，Beijing 100190，China；
3．Graduate school of Chinese academy of social sciences，Beijing 102488，China)

**Abstract:** The longest noun phrases carry abundant syntactic and semantic information, frequently corresponding to syntactic components; and play a certain semantic role in sentences. Recognition of the longest noun phrase plays an important role in Natural Language Processing and is the basis for analyzing and understanding sentence structure and semantics. By comparing the essence of the different longest noun phrases, this paper defines the fundamental concept of the longest noun phrase in Tibetan language based on the perspective of syntax tree. Total of 6038 sentences are extracted from the syntax tree corpus, the structure type, boundary feature and frequency of longest noun phrases are analyzed, and the longest noun phrases are recognized using the sequence annotation model and the syntactic analysis model. The correct rate, recall rate and F1 value of the recognition results of applying sequence annotation model are 87.14%, 84.72% and 85.92% respectively. The correct rate, recall rate and F1 value of the recognition results of applying syntactic analysis model are 87.66%, 87.63% and 87.65% respectively.

**Key words:** Tibetan syntax tree; the longest noun phrase; type of noun phrase

## 0 引言

人通过识别文本中的实体、概念来理解文本，理解了文本中的实体概念，在某种程度上就理解了文本的大致内容。名词或名词短语经常被用来

**收稿收稿日期**：2018-06-01；**定稿日期**：2018-07-12
**基金项目**：国家语委科研项目（ZDI135-17）；

表达实体、概念。名词或名词短语的识别在自然语言处理中占重要地位，是一个句子的主要组成部分，它携带着丰富的句法和语义信息，是分析和理解句子意义和结构的基础。在自然语言信息处理领域，名词短语的识别和结构分析正确可以提高机器翻译、信息检索、文本分类、自动句法分析等自然语言处理系统的性能。

在藏语信息处理领域，词法分析取得了丰富的成果[1-4]，信息处理逐渐从词法分析为主过渡到句法、语义和篇章分析为主的阶段。从句法分析的角度来看，研究内容表现在两个方面：一是句子的识别，二是句法分析。句子识别主要讨论如何从连续文本中切分出一个个句子。如从语言规则出发，可以根据藏语动词语尾的特点，构建句子边界切分标记库，实现句子切分[5-6]；或者采用规则和统计相结合的方法识别句子边界[8-12]；也有一些研究，在双语语料对齐研究中，探讨句子的边界问题[13-14]。句法分析主要讨论基于短语结构的句法分析[15]和基于依存语法的句法分析[16-17]。为了降低句法分析的难度，研究者倾向于采用组块分析方法进行局部句法分析，其中名词组块是组块分析的重要部分[18-20]。尽管局部句法分析取得了一定的成果，但是，从语言工程实践角度来看，成系统、上规模的藏语句法树库资源极其缺乏，实用的句法分析工具也未见公开。

本文开展基于藏语短语结构句法树库的最长名词短语研究，从构建短语结构树的角度，理清最长名词短语的定义、类别。从句法树库中选取了 6038 个句子，对名词短语的类型、结构等统计分析。初步构建藏语最长名词短语识别器，分析识别效果和存在的问题。

## 1 最长名词短语定义

台湾学者 Chen 研究英语名词短语的分类，总结出三种名词短语：最短名词短语、最长名词短语和普通名词短语。所谓最短名词短语是指不包含其他名词短语的名词短语，最长名词短语是指不被其他名词短语所包含的名词短语。普通名词短语是不具有任何限制的名词短语[21]。周强把名词短语也分成三类：最短、最长和一般名词短语。一般名词短语指所有不是最长和最短的名词短语[22]。两种分类类似，但内涵有差别，如在对待单个词构成短语时，前者的基本思想是，一个词可以构成最长名词短语；但后者认为，一个词构成的短语不是最长名词短语。钱小飞在总结各种名词短语定义之后，区分了最长名词短语和表层最长名词短语，从他列举的例子中，可以观察得出，所谓表层最长名词短语是指在句法树的子树中，包含的第一个层级的名词短语，非表层最长名词短语是指表层最长名词短语中不包含动词短语的嵌套名词短语[23]。

Koehn 和 Knight 从句法树的角度界定最长名词短语和介词短语，即给定一个句子 *S* 和它的句法分析树 *t*，名词和介词短语是句子 *S* 的子树 $t_i$，它至少包含一个名词，但不包含动词，不被更大的名词短语和介词短语所包含[24]。Koehn 和 Knight 对最长名词短语的界定基于句法树，这个定义比较符合本文基于短语结构树的藏语最长名词短语的定义，藏语最长名词短语基于句法分析树，更加注重名词短语及其他短语在句法分析树上的位置。参考前人的研究成果，结合藏语句法分析树的实际情况，本文把藏语最长名词短语界定为：

给定一个藏语句子的句法分析树 *S*，最长名词短语是 *S* 的子树 *t*，*t* 是名词短语但 *t* 的父节点及祖先节点都不是名词短语。

这个概念界定比较宽泛，从句法分析树看，自顶向下，第一个名词短语就是本文所指的最长名词短语。

最长名词可以由单个名词、代词、数词等构成。如图 1 所示，KP-SBJ-AGE 短语的子节点 RP（人称代词提升为名词性短语），KP-OBJ-TAR 短语的子节点 NP，VP 短语的子节点 NP 为最长名词短语。

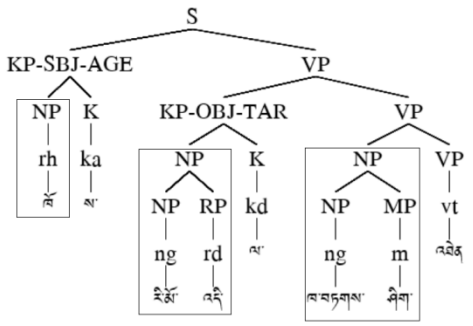



**图 1 句法树中的最长名词短语**

为了更细致地描述藏语最长名词短语，特做如下界定：

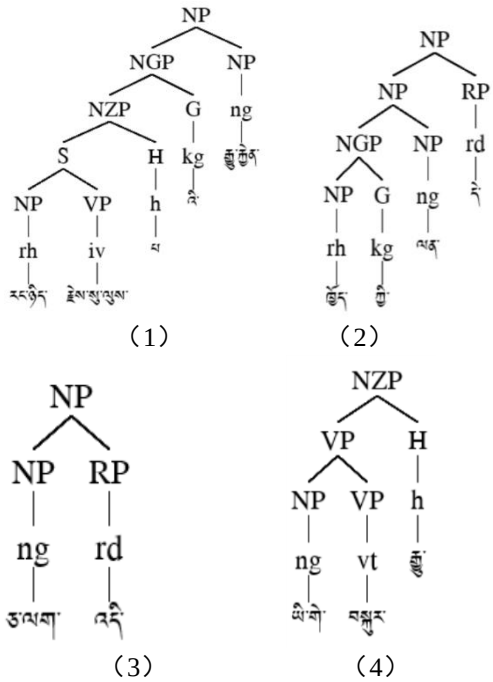
（1）最长名词短语是指中心词为名词的所有短语；最长名词短语的中心词位置可以居于短语首、短语中和短语末。如图 2（1）的中心名词居尾、图 2（2）的中心名词居中、图 2（3）的中心名词居首。

（2）最长名词短语可以由单个名词、代词、数词等构成；如图 2（3）中名词 ཅོ་ལག་ 和代词 འདི་

提升为短语,然后再与短语一起构成更大的短语。

（3）名词化标记可以作为最长名词短语的中心词,如图 2(4)名词短语的中心是名词化标记。

（4）最长名词短语可以是嵌套短语，包括内嵌名词化短语,如图 2(1)是嵌套的名词化短语，图 2（2）中嵌套有名词短语，图 2（4）内嵌套动词短语。




（1） （2）

（3） （4）

图 2 名词短语结构类型

## 2 句法树库及最长名词短语抽取

### 2.1 藏语句法树库介绍

藏语句法树库由中国社会科学院民族学与人类学研究所构建,句法分析采用了短语结构语法,本文研究材料来源于 1 万句基本句型句法分析树库。

在句法树库中，一个句子除了按照词切分之外，还包括词的词性信息、短语类型信息、句法功能信息、语义角色信息以及句子（或者结构）的关系信息。在短语类型层级的节点上，标注的信息包括短语类型、句法功能和语义角色。如果涉及到句子或者结构之间的关系，在短语的句法功能之后标注关系信息，例如：

((IP(S(KP-SBJ-AGE(NP ཨན་ད/ng )(K ས/ka ))(VP(KP-OBJ-TAR(NP(NGP (NP(NGP (NP ང/rh )(G འི/kg ))(NP ཚ་བ/ng ))(G འི/kg ))(NP ཅོ་/ng ))(K ལ/kd ))(VP(NP-OBJ(NP(NP ཨན་/ng )(ADJP གསར་ད/a ))(MP ཞིག/m ))(VP སྨན/vt ))))(PU །/xp )))（医生对我的胃病开了新药）

叶子节点（终结点）是词和词性。词与词性的上位节点是短语（非终结点），非终节点可以承载短语信息、句法功能信息、语义角色信息和句子关系信息。在上例中，KP-OBJ-TAR 表示带有格标记的名词短语（KP）的子节点在句子中充当间接宾语（OBJ），表示对象（TAR）语义角色。

基于短语结构语法的藏语句法树库标注符号可以分成三类：短语标注符号、句法标注符号和语义角色标注符号。

（1）短语标注符号包括 IP（带时体态的句子）、S（核心句）、NP（名词短语）、KP （带有格标记短语）、NZP（名词化短语）、VP（动词短语）、ADJP（形容词短语）、ADVP（副词短语）、ADZP（副词化短语）、NGP（领属关系短语）、VP（动词短语）、QP（量词短语）、MP（数词短语）、PRN（插入语短语）、IDE（独立成分）、UP（带助词标记短语）

（2）句法标注符号包括：SBJ（主语）、OBJ（宾语）、PRE（谓语）、ADV（状语）、APP（同位语）

（3）语义角色标注符号包括：AGE（施事）、PAT（受事）、TAR（对象）、DIR（方向）、SPA（处所）、TIM（时间）、MAN（方式）、INS（工具）、MAT（材料）、SOU（源点）、PUR（目的）、FAC（使役）、RES（结果）、BAS（依据）。

在句法树标注过程中还需要说明的一些标注符号包括：I（时体态）、T（时）、E（态）、H（名词化标记）、AUX（助动词）、G（连接标记-属格）、PL（复数标记）、U（助词标记）、Z（后缀标记）、RP（人称代词）、K（格标记）、Y（语气标记）。词性标注体系可以参阅《中国语言生活绿皮书 A006》[1]

### 2.2 最长名词短语抽取

为了研究最长名词短语的内部结构，展示藏语最长名词短语的特性，作者首先从句法树库中选择一定的句法树，抽取出最长名词短语。抽取方法主要是根据嵌套括号标记，找到句法树中最长的、节点标记类型为 NP 的短语，并将该节点的文本表示抽取出来；同时，将构成短语的每个词语的类别也抽取出来。例如：

[1] 赵小兵，孙媛，龙从军等：信息处理用现代藏语词类标记集规范(草案)，教育部语言文字信息管理司组编，《中国语言生活绿皮书 A006》，商务印书馆，2015 年 7 月。

((IP(S(KP-SBJ-AGE(RP(rh ཞྭ་ ))(K(ka ཀྱི་ )))(VP(NP(NP(NP(NP(ng རྒྱ་མཚོ་ ))(G(kg འི་ )))(NP(ng གནས་སྐབས་ )))(ADJP(a ཅི་ཙམ་ )))(VP(vt ཟེར་ ))))(PU(xp ། ))))

以 NP 节点为例，抽取出该节点对应的子树为(NP(NP(NP(NP(ng རྒྱ་མཚོ་ ))(G(kg འི་ )))(NP(ng གནས་སྐབས་ )))(ADJP(a ཅི་ཙམ་ )))，对应的名词短语为 NP，由(NP(ng རྒྱ་མཚོ་ ))(G(kg འི་ )))(NP(ng གནས་སྐབས་ ))构成。类似的，从该句子中还可以抽取出其他 KP 节点下的 RP（RP 等同于 NP）。由 RP 独立构成。根据该规则可以从句子中抽取出两个最长名词短语，分别是： （RP(rh ཞྭ་ ))和(NP(ng རྒྱ་མཚོ་ ))(G(kg འི་ )))(NP(ng གནས་སྐབས་ )))(ADJP(a ཅི་ཙམ་ )))。

## 3 最长名词短语结构分析

前文谈到，最长名词短语可以是独词构成，也可以由多词构成。通过了解最长名词短语在真实文本中的分布状况，有针对性地采取一些处理策略，有利于提高句法分析的精度。理论上，符合藏语语法规则的名词短语的长度可以无限递归增长，相应地，结构类型也会增加。根据对 6038 句句法树中的最长名词短语统计结果来看，最长的名词短语包含 18 个音节，且是内嵌名词化短语，如：དང་རྒྱུད་རིང་བའི་སློབ་ནས་སློབ་པ་སྡུག་སྡུད་ཞེས་པ་དེ་ཚོ་བློ་འགྲུག་ཅིང་ཡས་དེ་（耐心地教导那个有坏行为的小学生），短语内嵌一个小句，通过名词化标记 ཡས་ 名词化后，与指代词 དེ་ 构成名词短语。

经对不同结构类型的最长名词短语统计，统计数据如表格 1 所示，在语料中出现次数小于 10 次的结构类型，共计 450 种，超过 90%的结构类型出现总次数只占 9%。这些类型不是最长名词短语的强势组合模式。其中有 316 种结构只出现 1 次，如：ADJP+VP+H+G+NP+NP，གསར་པ་བརྩེ་བའི་མི་དམངས་སློབ་ཁང་（新建的人民医院）；NP+ADJP+ADVP+VP+AUX+H+RP，བོད་སྐད་ཡག་པོ་འདི་འདྲས་ གསུང་ཐུབ་ཡས་དེ་（藏话能说得这样好）等。有 85 种结构，每种出现 2 次，如：NP+ADVP+ADJP，དོ་སྣང་ཉ་ཅང་ཆེན་པོ་（非常大的注意力）。少于 10 次的名词短语结构类型出现次数如**表格 2** 所示。

**表格 1 低频最长名词短语结构类型的种类及出现次数**

| 类型总数 | 316 | 85 | 32 | 10 | 3 | 11 | 7 | 8 | 7 |
|---|---|---|---|---|---|---|---|---|---|
| 次数 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |

**表格 2 频次大于 10 的最长名词短语的结构类型及出现次数**

| 序号 | 类型 | 频次 | 实例 | 实例翻译 |
|---|---|---|---|---|
| 1 | NP | 4778 | རླུང་སྒྲ་ | 风声 |
| 2 | NP+RP | 640 | རི་མོ་འདི་ | 这幅画 |
| 3 | NP+G+NP | 588 | རྒྱ་མཚོའི་གནས་སྐབས་ | 海洋的情况 |
| 4 | NP+ADJP | 411 | སྐད་མཐོན་པོ | 高的声音 |
| 5 | NP+MP | 390 | ཁ་བཏགས་ཤིག་ | 一条哈达 |
| 6 | RP+G+NP | 373 | ཁྱོད་ཀྱི་དཔེ་ཆ་ | 你的书 |
| 7 | NP+NP | 264 | བོད་ཡིག་སློབ་འཕྲིང་ | 藏文中学 |
| 8 | RP+PL | 188 | ཁོ་ཚོ | 他们 |
| 9 | NP+ADJP+MP | 129 | སྨན་གསར་བ་ཞིག་ | 一种新药 |
| 10 | NP+RP+G+NP | 86 | རི་དེའི་དབྱིབས་ | 那山的形状 |
| 11 | NP+Z | 52 | ཟླ་བཟང་ལགས་ | 洛桑拉 |
| 12 | NP+G+NP+MP | 49 | དབྱིན་ཡིག་གི་རྒྱགས་ཚང་ཞིག་ | 一次英语小测验 |
| 13 | NP+PL | 49 | ཕྱིས་པ་ཚོ་ | 孩子们 |
| 14 | NP+G+NP+RP | 46 | དབྱིན་ཡིག་གི་ཚིག་འདི་དག་ | 这些英语的句子 |
| 15 | NP+G+NP+G+NP | 44 | ང་ཚོའི་རྐང་རྡེད་སྡུ་བའི་ཟུ་ཁག་ | 我们的足球队 |
| 16 | NP+VP+H+G+NP | 41 | སྐོབས་པ་སྣྲིད་པའི་ངང་ | 自豪的样子 |
| 17 | NP+ADJP+RP | 40 | ཙོག་ཙོ་སྐོར་སྐོར་དེ | 那圆圆的桌子 |
| 18 | RP+G+NP+G+NP | 40 | བའི་དུས་ཚོད་ཀྱི་རིན་ཐང་ | 我的表的价值 |
| 19 | NP+MP+G+NP | 39 | དུས་ཚོད་བཅུད་ཀྱི་ཆུང་ | 八点整 |
| 20 | NP+NP+MP | 38 | མེ་ཏོག་ཚ་པ་ཞིག་ | 一束花 |
| 21 | RP+G+NP+RP | 38 | ཁྱོད་ཀྱི་ལན་དེ་ | 你的那回答 |
| 22 | NP+NP+G+NP | 36 | ལུས་རྩལ་སྦྱོང་བརྡར་གྱི་རིགས་ | 体育锻炼的种类 |
| 23 | RP+MP | 35 | ཁྱེད་རང་གཉིས་ | 你们俩 |

| | | | | |
|---|---|---|---|---|
| 24 | NP+G+NP+ADJP | 31 | རྒྱ་མཚོའི་གནས་ཚུལ་ཇི་ཙམ་ | 海洋的情况怎样 |
| 25 | RP+NP | 30 | ང་གཟུགས་པོ་ | 我身体 |
| 26 | NP+QP+MP | 22 | ཁ་པར་ཟེངས་གཉིས་ | 两次电话 |
| 27 | RP+G+NP+NP | 22 | ཁྱོད་ཀྱི་ལེ་ཚན་རེའུ་མིག་ | 你的栏目表 |
| 28 | NP+K+VP+H+G+NP | 19 | རང་ཉིད་ཀྱིས་བརྩམས་པའི་རོལ་དབྱངས་ཤིག་ | 自己作的一首曲子 |
| 29 | NP+G+NP+NP | 17 | ཡེ་ཤུའི་འཁྲུངས་སྐར་ཉིན་ | 耶稣诞生的日子 |
| 30 | NP+ADJP+G+NP | 15 | རི་ཆུང་ཆུང་གི་རྩེ་ | 小山的山头 |
| 31 | NP+NP+RP | 14 | ཕྱག་ today་སྨད་ཁོག་དེ | 那后腔羊肉 |
| 32 | NP+NP+NP | 13 | ཟླ་ཏང་བོད་ཡིག་སློབ་འཕྲིང་ | 拉藏藏文中学 |
| 33 | VP+H+G+NP | 12 | ཉམ་སྐྱ་བཀྲེབ་པའི་ཁ་ལག | 发霉的食物 |
| 34 | NP+RP+PL | 11 | གང་སྡིགས་འདི་ཚོ་ | 这些垃圾 |
| 35 | RP+K+VP+H+G+NP+RP | 11 | ངས་བསྐུར་བའི་སྐུར་མ་དེ་ | 我寄送的包裹 |
| 36 | NP+MP+Z | 10 | དུས་ཚོད་བཞི་ཡས་མས་ | 四点钟左右 |
| 37 | NP+Z+G+NP | 10 | སྒོལ་དཀར་ལགས་ཀྱི་གཟིམ་ཁང་ | 卓嘎拉的寝室 |
| 38 | RP+G+NP+ADJP | 10 | ངའི་མཛའ་གྲོགས་མང་པོ་ | 我的许多好友 |
| 39 | RP+PL+G+NP | 10 | ཁྱེད་རང་ཚའི་ཁ་ཡུལ་ | 你们的家乡 |

实际上，出现频次最高的前 10 个约占全部最长名词短语的 87%。尤其是单个名词和代词充当的短语占比高于 64%。频次较高的前 10 种类型结构都不包含嵌套名词化短语，长度也不大，最多由四个音节构成，详细情况如表格 2 所示。

从表格 2 中可以归纳如下几种类型：

**（1）独词短语** 包括名词、代词、数词都可以直接构成独词短语，RP，NP，MP，例如：

((IP(IP(S(RP(rh ང་))(VP(KP-ADV-SPA(NP(NP(ng ཟླ་ཐང་))(NP(NP(ng བོད་ཡིག་))(NP(ng སློབ་འཕྲིང་ ))))(K(kx ཏུ་)))(VP(vi འགྲོ་ )))(I(T(h ཀེ་))(E(ve ཡོད་ ))))(PU(xp ། ))))，

((IP(S(NP(ng ཟླ་བཟང་ ))(VP(KP-ADV-SOU(NP(ng འཁྲིན་བཙུར་))(K(kc ལས་)))(VP(vi རྒྱལ་)))) (PU (xp ། ))))

**（2）独词加标记（复数、敬语和约数标记）** 名词、代词带复数、敬语标记构成 RP+PL，NP+PL，NP+Z，数词可以带约数标记构成 MP+Z，例如：

((IP(IP(S(NP(NP(ng ཕྱི་ལ་ད་ ))(PL(pl ཚོ་ )))(VP(ADZP(VP(iv ཟླལ་བསྒྱིགས་ ))(U(c ནས་ )))(VP(vi བཙུད་ ))))(E(ve ཡོད་ )))(PU(xp ། ))))

((IP(S(KP-SJ-AGE(NP(RP(rh ཁོ་))(PL(pl ཚོ)))(K(ka ས་))) (VP(ADZP(VP(NP(ng འདེམས་ཤོག་ ))(VP(vt འཐབས་))) (U(c ནས་ )))(VP(NP(ng ཀློག་ཁེ་ ))(VP(vt བདམས་ )))))(PU(xp ། ))))

**（3）双词短语** 根据中心词的位置不同可以分成：中心词居后和中心词居前，前者构成的类型是 NP+NP，后者构成的类型有 NP+RP、RP+MP、NP+MP、NP+ADJP，例如：

((IP(S(NP(NP(ng ཆངས་འཚོར་ ))(MP(m ཞིག་ )))(VP(KP-ADV-SOU(NP(NP(ng ཀུ་ལམ་ ))(NP(ng སྐྱེང་ )))(K(kc ནས་ )))(VP(vi སོང་ ))))(PU(xp ། ))))

((IP(S(NP(NGP(NP(NP(ng རྒྱ་ཐེན་ ))(RP(rd དེ་ )))(G(kg འི་ )))(NP(ng གཉིང་ )))(ADJP(NEG(dn མེ་ ))(ADJP(a ཟབ་ ))))(PU(xp ། ))))

((IP(S(NP-SBJ-TOP(RP(rh ཁོ་ ))(U(up ཡིས་ )))(VP(NP(NP(ng གཟུགས་ཐང་ ))(MP(m ཞིག་ )))(VP(vl རེད་ ))))(PU(xp ། ))))

**（4）三词短语** 根据中心词的位置不同可以分成：中心词居后和中心词居前，前者构成类型有：NP+G+NP、RP+G+NP、NP+NP+NP、NP+VP+H[2]，后者构成类型 NP+ADJP+MP、NP+QP+MP，例如：

((IP(S(KP-SBJ-POS(NP(NGP(NP(ng ཁང་བ ))(G(kg འི་ )))(NP(ng ཐུང་ )))(K(kp ན་ )))(VP(NP(NP (ng བྱ་ག་ ))(MP(m ཞིག་ )))(VP(ve ཡོད་ ))))(PU(xp ། ))))

((IP(IP(S(RP(rh ང་ ))(VP(KP-ADV-SPA(NP(NP(ng ཟླ་ཐང་ ))(NP(NP(ng བོད་ཡིག་ ))(NP(ng སློབ་འཕྲིང་ ))))(K(kx ཏུ་ )))(VP(vi འགྲོ་ ))))(I(T(h ཀེ་ ))(E(ve ཡོད་ ))))(PU(xp ། ))))

((IP(S(KP-SBJ-AGE(RP(rh ང ))(K(ka ས་ )))(VP(NZP(VP(NP(ng ཁེན་ཁ་ ))(VP(ve ཡོད་ )))(H(h ཡ་ )))(VP(vt ཚམ་ ))))(PU(xp ། ))))

其他类型的短语都是在上述四种类型的基础上扩充，本文不再一一阐述。

藏语最长名词短语的边界词也具有明显特征。名词短语经常添加格标记，格标记是名词短语最重要的右边界特征词之一，还有包括数词、指示代词、复数标记、敬语标记、形容词等边界特征

[2] 名词化短语不作为修饰语时，名词化标记是短语的中心。

词。从本文数据统计结果看，作为名词短语一部分的、典型右边界词中，数词有 1313 个，复数标记 267，代词的 905，不作为名词短语一部分的右边界特征词主要是格标记，共有 4752 个名词短语有格标记。名词短语左边界特征词不明显，判断难度相对大一些。

## 4 最长名词短语识别实验

本文使用两种方法进行最长名词短语识别实验：一种采用序列标注方法，把名词短语识别转换为对名词短语边界特征词的识别；另一种采用句法分析方法，在整个句法树生成过程中，统计名词短语子树分析的结果。

在实验中，共使用 6038 句藏文句法树进行实验，将其中 5000 句作为训练语料，其余 1038 句作为测试语料，其实验结果如表格 3 所示。

表格 3 短语识别情况

| 实验方法 | 抽取数 | 实有数 | 正确数 | 正确率 | 召回率 | F 值 |
|---|---|---|---|---|---|---|
| 基于句法分析的方法 | 2290 | 2304 | 1947 | 85.02% | 84.51% | 84.76% |
| 基于序列标注的方法 | 2240 | 2304 | 1952 | 87.14% | 84.72% | 85.92% |

1、基于句法分析的方法：使用伯克利大学的 Berkeley Parser 在训练集上训练一个句法分析器，对测试语料进行句法分析，提取其中的最长名词短语。句法分析完全正确的句子比例为 32.49%。从测试语料中共识别出短语 2290 个，其中 1947 个是测试语料中实际有的短语，测试语料中实有名词短语的总数为 2304，名词短语识别的正确率、召回率和 F1 值分别为 85.02%、84.51%、84.76%。

2、基于序列标注的模型：将最长名词短语识别转化为序列标注问题，根据词语在名词短语中的位置，给其分别赋予位置标签，本文采用常用的 BMESO 标签集。使用 CRF++进行序列标注的训练和预测。从测试语料中共识别出短语 2240 个，其中 1952 个是测试语料中实际有的短语，测试语料中实有名词短语的总数为 2304，名词短语识别的正确率、召回率和 F1 值分别为 87.14%、84.72%、85.92%。

从表中数据可以看出，在识别最长名词短语任务中，基于序列标注的方法要比基于句法分析的方法稍好。

在基于句法分析方法中，缺乏格标记的名词短语容易出错，尤其是 VP 的孩子节点，通常，VP 可以分析为 NP 和 VP，也可以是 ADVP 和 VP，从训练语料的情况来看，分析为 NP 和 VP 的概率相对较大，因此，模型在预测时经常会把 ADVP 预测为 NP。例如图 3 所示。

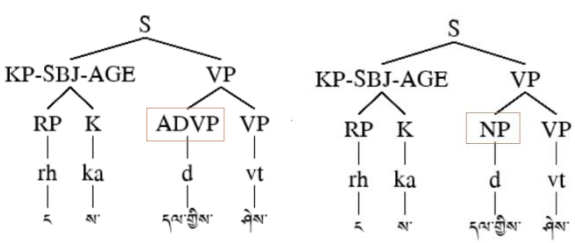



图 3 句法分析模型预测结果（右），标准答案（左）

VP 节点的孩子节点应该分析为 ADVP 和 VP，实验结果则分析为 NP 和 VP，其原因是，在训练语料中，VP 孩子节点分析为 NP 和 VP 结构的概率要远远大于 ADVP 和 VP 结构。但是这种错误应该比较容易纠正，དག་ཐྱོས 已经被标注为副词性短语，在词层级有标记 d，这已经说明它不可能是名词或名词性短语，针对这种错误，可以通过一致性检测处理进行纠正。

在基于序列标注模型的分析方法中，名词短语的长度大小会影响分析结果。太长的、具有嵌套结构的名词短语经常会被“切碎”，例如：བོད་སྐད་ཡག་པོ་འདི་འདྲས་གསུང་ཐུབ་ཡས་དེ་，序列标注结果为[བོད་སྐད་ཡག་པོ་འདི་འདྲས་][གསུང་ཐུབ་ཡས་དེ་]，实际上整个字串是一个短语。这是序列标注在处理长距离边界识别问题中普遍存在的问题。

在句法分析中，首先经过分词和词性标注过程，这个过程中的错误也直接会导致两种模型对最长名词短语的识别错误。例如：

ཚོ་ཆོས་ཁང་བའི་ཟླད་ཞིག་གསོ་བྱས།

两种模型都分析错误，在分词阶段 ཞིག་གསོ་བྱས 被切分为 ཞིག་/གསོ་/བྱས/，而 ཞིག་ 是名词短语右边界的典型特征词，因此两种模型都把 ཞིག་ 识别为名词短语的右边界。本句正确切分和识别结果应该为：

((IP(S(KP(NP(NP(rh ཚ་ ))(PL(pl ཙ )))(K(ka ས་ )))(VP(NP(NGP(NP(ng ཁང་བ ))(G(kg ཨི་ )))(NP(ng ཟླད་ )))(VP(vt ཞིག་གསོ་བྱས ))))(PU(xp ། ))))

本文实验以基本句型语料为主，从前文的统计分析也可以看出，较长的名词短语所占比例不大，因此在实验中，基于序列标注模型的处理结果要好于句法分析模型。

## 5 结语

最长名词短语识别是句法分析的一项重要子任务，本文在藏语句法树库建设中，针对最长名词短语问题，从句法树角度界定了最长名词短语的定义，专门分析了最长名词短语的结构类型，并采用句法分析方法和序列标注方法分别进行实验，考察最长名词短语的识别结果，从实验结果来看，在针对小规模语料实验中，序列标注的方法比句法分析的方法稍好。但是，本结果也许与实验的语料类型有关，序列标注对短距离标注任务效果明显，从最长名词短语结构分析来看，本次语料对序列标注模型有利。由于受到语料规模和句法分析文法类型的限制，本文未能开展基于神经网络的句法分析实验，这是今后努力的方向。藏语句法分析急需要在两个方面开展工作：扩充句法树库规模；完成短语结构树与依存句法树库之间的转换，这两个问题是我们近期研究的重点任务。

## 参考文献


[1] 李博涵,刘汇丹,龙从军,吴健.基于深度学习的藏文分词方法[J].计算机工程与设计,2018,39(01):194-198.

[2] 李亚超,江静,加羊吉,于洪志.TIP-LAS:一个开源的藏文分词词性标注系统[J].中文信息学报,2015,29(06):203-207.

[3] 刘汇丹,诺明花,赵维纳,吴健,贺也平.SegT:一个实用的藏文分词系统[J].中文信息学报,2012,26(01):97-103.

[4] 史晓东,卢亚军.央金藏文分词系统[J].中文信息学报,2011,25(04):54-56.

[5] 赵维纳. 基于法律文本的藏语句子边界识别，北京语言大学博士论文，2012 年 6 月。

[6] 赵维纳,于新,刘汇丹,李琳,王磊,吴健. 现代藏语助动词结尾句子边界识别方法[J]. 中文信息学报,2013,(01):115-119.

[7] 赵维纳. 基于法律文本的藏语句子边界识别[C]//第五届全国青年计算语言学研讨会论文集(C).中国中文信息学会,2010:7.

[8] 李响,才藏太,姜文斌,吕雅娟,刘群. 最大熵和规则相结合的藏文句子边界识别方法[J]. 中文信息学报,2011,(04):39-44.

[9] 才藏太. 基于最大熵分类器的藏文句子边界自动识别方法研究[J]. 计算机工程与科学,2012,(06):187-190.

[10] 徐涛,加羊吉,于洪志. 统计与规则相结合的藏文句子自动断句方法[J]. 云南大学学报(自然科学版),2012,(06):653-657+663.

[11] 马伟珍,完么扎西,尼玛扎西. 藏语句子边界识别方法[J].西藏大学学报(自然科学版),2012,(02):70-76.

[12] 仁青吉,安见才让. 藏文句子边界自动识别方法的研究[J]. 信息与电脑(理论版),2014,(08):62-63.

[13] 于新,吴健,洪锦玲. 基于词典的汉藏句子对齐研究与实现[J]. 中文信息学报,2011,(04):57-62.

[14] 华却才让. 藏汉句子局部对齐策略的研究[J]. 青海师范大学学报(自然科学版), 2010, (04):39-43.

[15] 扎西加. 上下文无关文法与藏语句法分析[J].西藏大学学报[自然科学版）, 2013, 28(2):37-42.

[16] 扎西加,多拉. 藏语依存树库构建的理论与方法探析[J].西藏大学学报[自然科学版）, 2015, 30(2):76-83.

[17] 华却才让,赵海兴. 基于判别式的藏语依存句法分析[J].计算机工程,2013,39(4):300-304.

[18] 江荻.现代藏语组块分词的方法与过程[J].民族语文,2003(04):30-39.

[19] 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(06):165-168.

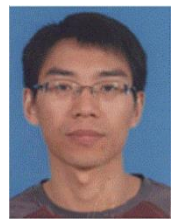
[20] 王天航,史树敏,龙从军,黄河燕,李琳.基于错误驱动学习策略的藏语句法功能组块边界识别[J].中文信息学报,2014,28(05):170-175+191.

[21] Kuang-hua Chen and Hsin-Hsi Chen：Extracting Noun Phrases from Large-Scale Texts:A Hybrid Approach and Its Automatic Evaluation[C]// Proceedings of the 32nd ACL Annual Meeting, 1994, pp. 234-241.

[22] 周强,孙茂松,黄昌宁.汉语最长名词短语的自动识别.软件学报, 2000, 11( 2): 195-201.

[23] 钱小飞,侯敏.面向信息处理的汉语最长名词短语界定研究.语言文字应用.2017,2:127-134

[24] Philipp Koehn, Kevin Knight. Feature-Rich Statistical Translation of Noun Phrases[C]//Proceedings of the 41st Annual Meeting of the Association, for Computational Linguistics, July 2003, pp. 311-318.


龙从军（1978—），博士，副研究员，主要研究领域为藏语计算语言学。
E-mail：longcj@cass.org.cn



刘汇丹（1982-），博士，副研究员，主要研究领域为操作系统中文信息处理、多语言信息处理。
E-mail：huidan@iscas.ac.cn



周毛克（1993—），硕士研究生，主要研究领域为藏语自然语言处理。
E-mail：zmk_muc@163.com