基于词对依存分类的藏语树库半自动构建研究

华却才让[13] 姜文斌[2] 赵海兴[1] 刘群[2]

(1.青海师范大学藏文信息研究中心, 青海 西宁 810008;

2.中国科学院计算技术研究所智能信息处理重点实验室，北京 100190;

3.陕西师范大学计算机学院, 陕西 西安 710062.)

摘要：依据依存句法理论，本文制订了藏语句法标注体系及层次结构。通过分析构建藏语依存树库中存在的问题，提出了半自动依存树库构建模式，针对藏语特性提出了融合丰富特征的词对依存分类模型和依存边标注模型，实现了依存树库构建可视化工具，校对构建了 1.1 万句藏语依存句法树后，在基线系统下经实验验证，依存识别正确率提高了 3%，大大提高了构建藏语依存树库工作的进展。

关键词：藏语依存句法；词对依存分类；藏语树库；藏语依存标注工具

中图分类号：TP391 **文献标识码：**A

Semi-Automatic Building Tibetan Treebank based on Word-Pair Dependency Classification

HUA Quecairang[13]，JIANG Wenbing[2]，Haixing Zhao[1]，LIU Qun[2]

1.Qinghai Normal University Tibetan Information Research Center, Xining, 810008; 2.Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190; 3.Computer Science School of Shaanxi Normal University, Xi'an 710062.

Abstract: According dependency syntactic theory this paper gave Tibetan typed dependencies and its hierarchy, and then we analyzed some problems in building Tibetan dependency Treebank. We proposed a mode to construct dependency tree semi–automatically, it includes word-pairs dependency classification model and dependency edges annotation model with rich features template based on Tibetan language grammar. And we implemented visualized tool which used to build and proofreading 11 thousand sentences Treebank. On the baseline system the experimental results show that, the dependency recognition accuracy obtains an improvement of 3%.

Key words: Tibetan dependency syntax; word-pair dependency classification; Tibetan Treebank; Tibetan dependency annotation tool

1. 引言

依存句法树库作为依存句法分析、句法机器翻、文本挖掘等热门研究领域的支撑语料，其重要性不言而喻。藏语依存句法树构建，从句法标注规范、句法树库构建及其规模均比较滞后。文献[1]结合纯手工构建的藏语依存句法树库(Tibetan Dependency Treebank, TDT)规模为 1 万句左右，采用一层感知机判别式方法训练模型，在 3 百句测试集上的依存识别正确率达到 81%[1]，中心词识别正确率为 87%，完整依存标注句子正确率为 34%。藏语作为 SOV 语序结构，并含有丰富的格助词接续规则，一句子中中心词在句末，直接宾语和间接宾语在主谓之间[2]，导致加大了纯手工标注依存句法或修改小规模训练语料上训练模型，解码分析得到正确率不算高的句法树。因此本文提出基于依存词对分类的藏语依存树库半自动构建方法，在分词标注的语料和句法分析器分析得到的句法树(标注结果不完全正确)上，呈现出比较直观和具有辅助提示功能的依存标注和修改功能。一定程度上加快了藏语依存句法的标注进展，保证了依存标注的正确性。并利用此方法对已有藏语依存句法树库的修改和补充，对已有藏语依存句法分析的正确率提升了 3%。

基金项目：本课题得到国家自然科学基金项目(61063033, 61163018) , 教育部“春晖计划”合作科研项目(Z2012102)资助. **第一作者简介**：华却才让(1976-),男, 副教授, 博士生, 主要研究方向为藏语词法分析、句法分析和机器翻译.

本文第 2 部分介绍了藏语依存句法、标注体系和树库构建情况；第 3 部分提出了依存词对分类模型和特征抽取以及半自动依存辅助模式，包括依存关系辅助模式和依存边标注辅助模型；第 4 部分讨论了基于依存词对分类的半自动句法树库构建及句法分析实验结果。

2. 句法理论

2.1 藏语依存句法

句子中词与词之间存在的支配和被支配的关系称作依存关系。其理论可追溯到上世纪五十年代，Tesniere.L 在其著作《结构句法基础》中首次提出来“依存语法”的概念，主张每个句子有一个中心词(一般为动词)，支配句子中的其它成分，而它本身不受任何句子成分的支配。此后为总结依存句法的概念 1970 年 Robinson J.J 提出了句法依存关系的四条公里，称之为依存语法的理论基础[3]，描述为：（1）一个句子中只有一个成分是核心成分；（2）其它成分直接依存于某一成分；（3）任何成分后不能依存于两个或两个以上的成分；（4）如果成分 P 依存于成分 Q，那么 P 和 Q 之间的任意成分 R 就不能依存到 P 和 Q 所构成的跨度之外，成分 R 或者直接依存于 P，或者直接依存于 Q，或者直接依存于 P 和 Q 之间的某一成分，总之，依存成分之间不能交叉。

依据现有其他语言的依存树库构建经验，如斯坦福大学英语依存标准[4]和文献[5]中阐述的中文依存体系，在设计藏语依存句法关系时，同样考虑了对依存关系的易于理解和使用的要求，使其能为藏语句法研究提供真实文本标注素材，便于语言学家从中总结语言规则和规律。用一个三元组 r(v,w)描述词对之间依存关系，意为词 v 支配词 w，v 是主词，w 是从词，r 称作 v 向 w 的下依存关系。如 SUB(བཤད/vt-9,སློབ་དཔོན་/nr-1)可解释为“བཤད/vt”的主语是“སློབ་དཔོན་/nr”。这种句法依存关系有利于从句子的词语线性排列中获取内部语言信息。如藏语句子“སློབ་དཔོན་གྱིས་སྒོག་ཐོག་རྣམས་ལ་སྡང་ས་འདུག་གྱི་རྣམ་ཐར་བཤད།(卓嘎给同学们讲朗萨雯蚌传。)”的依存句法关系可描述为如下：

SUB(བཤད/vt-9,སློབ་དཔོན་/nr-1) GZC-X(བཤད/vt-9,གྱིས/gx-2)

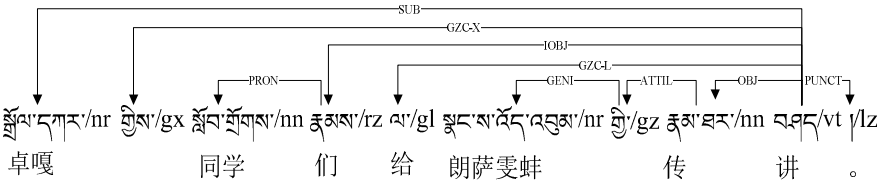
PRON(རྣམས་/rz-4,སྒོག་ཐོག་/nn-3) IOBJ (བཤད/vt-9,རྣམས་/rz-4)

GZC-L(བཤད/vt-9,ལ་/gl-5) GENI(གྱི་/gz-7,སྡང་ས་འདུག་/nr-6)

ATTIL(རྣམ་ཐར་/nn-8,གྱི་/gz-7) OBJ(བཤད/vt-9,རྣམ་ཐར་/nn-8)

PUNCT(བཤད/vt-9,།/lz-10)

依存关系可转换为有向图句法树 (V,E) 模式，其中每一个词表示图中的一个结点 $v\in V$，词对间的依存关系 r 为有向边 $e\in E$ 的标签。上面藏语句子依存分析树同样可以表示为图 1。



正确翻译：卓嘎 给 同学 们 讲 朗萨雯蚌 传 。

图 1 有向图模式藏语依存句法分析树

在没有边标签的依存句法分析应用中，我们可以藏语的依存树结构表示为如下形式：((སློབ་དཔོན་/nr) (གྱིས/gx) ((སྒོག་ཐོག་/nn) རྣམས/rz) (ལ་/gl) (((སྡང་ས་འདུག་/nr) གྱི་/gz) རྣམ་ཐར་/nn) བཤད/vt (།/lz))

2.2 句法标注体系及层次结构

传统藏语文法包括两部分，一是文法根本三十颂，讲藏语拼写结构、格助词和各类虚词的用法；二是字形组织法，主要讲以动词为中心的形态变化、时态，施受和能所关系等[6]。藏语句子是格助词、虚词和动词等依据句法理论发生结构关系而成的词语线性排列，其基本语序结构为<主语>+<间接宾语>+<直接宾语>+<结果补语>+<状语>+谓语+<状态补语>。据藏语词频统计，藏语语料库中格助词类的频率最高，其通用度稳定，句子中内部组织成分的层次结构也基本与格助词结合相关。例如句子中出现主格时，其前面一般为使动成分而后面部分为被动成分，即基本可确定主语和宾语。考虑到依存关系过于细致和庞大，会导致交互式句法分析器的鲁棒性、可操作性下降和统计数据的稀疏等问题，本文结合藏语实际切分标注语料制订了 36 类藏语依存关系[7]，其数量相对英语和汉语比较少，但尽可能使制订的句法依存关系满足藏语的各种语法现象，涵盖句法规律。句子中一条依存句法关系表示了一个支配词(一头结点)和一个被支配词(一个孩子结点)的二元支配关系。由于论文篇幅，本文中对依存句法关系的定义以及与其他主流依存句法体系的差异不作具体赘述，将在藏语依存句法体系一文中作详细讨论。藏语依存句法层次结构见表 1 所示：

表 1 藏语依存句法层次结构见表

ROOT - 谓语依存
AUX - 助动词
AUXIW - 助动词结构(还包括存在、判断、时态和祈使助词)
COP - 表语结构(COPULA)
ZHB - 状态补语结构
ARG - 论元
SUB - 主语
CSUB - 从句主语结构
ATTI - 定语(ATTIL|R 后置形容词有修饰成分)
GZC - 主|于|从格结构
OBJ - 宾语结构
DOBJ - 直接宾语结构
IOBJ - 间接宾语结构
POBJ - 介词宾语(FOBJ - 方位宾语结构)
AOBJ - 宾语提前结构
SQW - 特殊疑问词结构
SCOM - 结果补语结构
CLAUSE - 从句结构
COOR - 并列结构
CONJ - 连词结构
SUW - 离合词结构
SC - 待述词结构
MOD - 修饰
QUANT - 限定(限量、时间)结构
AFFIX - 词缀结构
QDIG - 数词附加语结构
GENI - 属格结构
YY - 陈述词结构(“ཏུ”)
OS - 饰集词结构
PRON - 代词结构

EXIAU - 存在助词结构(“ཡིན་གྱི་ཡོད་”)
DUP - 词的重叠结构
ADVMOD - 状语结构
ADVANG - 状态修饰结构(“ཅང་”)
APPOS - 同位语结构
GLC - 关联词结构
EXCL - 叹词结构
FASQP - 终结和疑问结构
PUNCT - 标点符号
DEP - 未知依存结构

2.3 藏语依存树库及问题

藏语依存树库的构建目前处在起步阶段，是一项比较庞大的工程。以现有依存库TDTreebank 1.0作为训练语料的统计句法分析器分析能力还比较弱，主要原因可归结为句法树库的规模小；人工标注的平均句子词数小于17[1]，对一些歧义结构，特别是复杂句子的的分析错误还很多。有了之前 TDTreebank 1.0 树库语料，目前藏语依存树库的构建大致分为三步，一为预处理，主要对生语料做断句、分词和词性标注，以及语料全半角等的统一工作。本文中使用了青海师范大学的藏语分词和词性标注规范[8]；二是机器分析，用句法分析器分析生成句法分析树；三是人工校对。其中第三步是比较枯燥的工作，但也是必须要做的工作，也是本文的主要研究介入点。为获得最佳的整体处理效果，人工校对时，需要提供词对依存关系辅助提示、比较可能的复杂歧义结构、交互式鼠标点击连接、修改支配关系、避免交叉依存关系标注以及长句依存标注支持等半自动功能。这些功能对人工校对或者纯人工标注具有事半功倍的效果。接下来为解决以上问题，我们在第3部分提出了词对依存分类模型支持的藏语依存树库构建方法，并实现了依存标注和修改工具TibetanDepBuilder V2.4。

3. 半自动句法标注

3.1 词对分类模型

3.1.1 模型

若给定包含N个词的一个句子，任意两个词之间都可能存在依存关系，寻找最可能的依存树的任务是从N*(N-1)种可能的依存边（无自环）完全图中寻找分数最大的树，于是，若y为句子x的依存树，则$(i,j)\in y$，$1\le i,j\le|x|$且$i\neq j$，其中(i,j)表示句子x中的词x_i和x_j之间存在有向边，x_i为词x_j的父节点；根据Eisner(1996)分解法可以将依存树y的分数$S(x,y)$表示为[9]：

$$s(x,y)=\sum_{(i,j)\in y}C(i,j)=\sum_{(i,j)\in y}(\sum\nolimits_k w_k\cdot f_k(i,j)) \qquad (1)$$

其中$f_k(i,j)$是依存词对i和j之间的第k个特征向量，w_k是该特征向量对应的参数向量，可通过训练样本获得。那么最大生成树模型[10]可以表示为：

$$\begin{aligned}\tilde{y}&=\arg\max s(x,y)\\&=\sum_{(i,j)\in y}C(i,j)\end{aligned} \qquad (2)$$

若候选词对依存分类的权重分数$C(i,j)$转换为概率模型$C(i,j)=P,\ (0\le P\le 1)$，概

率P表示候选依存边的强弱，那么基于概率的最大生成树模型可表示为(3)式, 表示对句子*x*解码生成的句法树集合中当前句法树*y*，连乘树中所有候选依存词对的概率值，最后获取概率最大的句法树：

$$\begin{aligned}\tilde{y} &= \arg\max s(x,y) \\ &= \arg\max_{y} \prod_{(i,j)\in y} C(i,j)\end{aligned} \qquad (3)$$

词对分类模型的任务是判断任意候选词对之间是否存在依存边，为有效获得词对依存分类概率值$C(i,j)$，本文采用最大熵分类器训练依存词对特征的概率值， w 是 ME 模型训练得到的参数向量，与每个特征向量是否对依存边有无贡献一一对应，表示贡献程度。$f(i,j,r)$ 是依存词对 i 和 j 之间的特征向量，表示该词对之间存在一个关系 r，其中 $r\in\{+,-\}$, 当r=+表示特征向量对该词对的依存边具有贡献，而r=-时却相反，如果一个特征 $f_k(i,j,r)\in f(i,j,r)$，则其值等于 1，表示该特征在训练语料中抽到的特征集中存在，否则不存在，那么词对的依存分类模型可定义为：

$$\begin{aligned}C(i,j) &= \frac{\exp(w\cdot f(i,j,+))}{\sum_r \exp(w\cdot f(i,j,r))} \\ &= \frac{\exp(\sum_k w_k\cdot f_k(i,j,+))}{\sum_r \exp(\sum_k w_k\cdot f_k(i,j,r))}\end{aligned} \qquad (4)$$

3.1.2 特征抽取

由于词对依存分类特征从一定程度上体现了语言学知识，其特征模板的设计和选择同样是影响机器学习的性能的主要因素之一，在最大生成树模型(McDonaldet al., 2005a)中提出了每个特征是由词 *i* 和 *j* 及前后的词语和词性构成[11]。为丰富句法特征信息，Collins distance (Collins,1996)方法提出了词 *i* 和 *j* 之间的距离句法信息[12]。这种方法解决了两个词之间顺序位置、相邻关系、是否动词居中以及两个词中间或左右是否存在标点符号等问题。藏语词对依存分类训练和解码中合成了以上两种句法特征生成模板。此外，在此基础上增加了以下特征：

1)两个依存词对 *i* 和 *j* 之间是否存在楔形分隔符：由于藏语句子中用楔形符号“།”表示复合句子句、同位语、从句结尾以及连词“ཏེ་”表示分隔符，类似于逗号和顿号功能。

2) 两个依存词对 *i* 和 *j* 之间是否存在主格：主格位于主语之后，表示主语为使动者，而中心词是一个及物动词。

3) 两个依存词对 *i* 和 *j* 之间是否存在于格：于格一般位于间接宾语和直接宾语之间，或者介词宾语末端，充当介词成分。

本文为藏语依存句法分析分别设计了 62 个藏语词对依存分类特征模板，63 个藏语词对依边标注特征模板，具体用于模型训练的特征模板见表 2 所示。

藏语词对依存分类特征模板内容分四类：(1).一元特征：定义为父结点或子结点(单个词)的特征信息构成；(2).二元特征：由父子结点共同的特征信息构成；(3).词对左右词性特征：考虑到更好地抽取到藏语格助词的搭配规律而补充了此特征信息；(4).距离特征：词对间包括其他词(结点)时的依存关系的特征信息。表 2 中 *p-word* 表示依存树中父结点词，*p-pos* 表示父结点的词性，*c-word* 表示依存树中子结点词，*c-pos* 表示子结点的词性，*p-pos-1* 表示父

结点左边的词性，*c-pos+1* 表示结点右边的词性，*d**表示词对间所包含其他词(依存结点)个数，当 *d**的值为负数时表示句法树中抽出词对的父结点在子结点的左侧，而当 *d**的值为整数时表示句法树中抽出词对的父结点在子结点的右侧。 藏语词对依存边标注特征模板分五类：除了四类词对依存分类特征模板，还有(5).边标注特征：*P-frame* 用于词对依存分类边标注时用的扩展特征，表示父节点的依存边信息。

表 2 藏语词对依存分类和边标注特征模板

特征分类	特征		
一元特征	*p-word, p-pos*	*p-word*	*p-pos*
	c-word, c-pos	*c-word*	*c-pos*
二元特征	*p-word, p-pos, c-word, c-pos*	*p-word, p-pos, c-word*	*p-word, p-pos, c-pos*
	p-word, c-word, c-pos	*p-pos, c-word, c-pos*	*p-word, c-word*
	p-pos, c-pos	*p-pos, c-word*	*p-word, c-pos*
词对左右词性特征	*p-pos-1, p-pos, c-pos-1, c-pos*	*p-pos, p-pos+1, c-pos-1*	*p-pos-1, c-pos+1, c-pos*
	p-pos-1,p-pos,c-pos,c-pos+1	*p-pos-1,p-pos, c-pos-1*	*p-pos-1,p-pos, c-pos+1*
	p-pos, p-pos+1, c-pos, c-pos+1	*p-pos, p-pos+1, c-pos+1*	*p-pos-1, c-pos-1, c-pos*
	p-pos, p-pos+1, c-pos-1, c-pos	*p-pos+1, c-pos, c-pos-1*	*p-pos+1, c-pos, c-pos+1*
	p-pos,c-pos,c-pos-1	*p-pos,c-pos,c-pos+1*	*p-pos, p-pos-1, c-pos*
词对距离特征	*p-word, d**	*p-pos, d**	*p-word, p-pos, d**
	*c-word, d**	*c-pos, d**	*c-word, c-pos, d**
	*p-word, p-pos, c-word, c-pos, d**	*p-word, p-pos, c-word, d**	*p-word, p-pos, c-pos, d**
	*p-word, c-word, c-pos, d**	*p-pos, c-word, c-pos, d**	*p-word, c-word, d**
	*p-pos, c-pos, d**	*p-word, c-pos , d**	*p-pos, c-word, d**
边标注特征	*P-frame*		

3. 2 半自动辅助模式

3. 2. 1 词对依存分类辅助提示

有了词对依存分类训练模型，接下来的工作是如何应用于一个词性标注好的句子中词语之间的依存标注，并即时呈现出辅助提示功能，一般标注依存词对时有两种方式，自底向上、自顶向下，其中自底向上为首先选择某个被支配词然后找出其所有可能的支配词；而自顶向下是首先在句子中选择一个支配词，然后找出所有可能的被支配词，也就是说自动给出对其余词语的被支配强弱的自动提示。本文采用了第二种自顶向下的模式。如图 3 所示。图中当前用鼠标选择的支配词(中心词)为“བཤད”，按词序号，第 8 和 9 已被选择为被支配词，8 为直接宾语，9 为楔形结束符。从词对依存分类辅助提示看出，剩余待被连接的 1、2、4 和 5 号词是最可能的被支配词，实际这些词分别为句子中的主语、主格、间接宾语和于格。

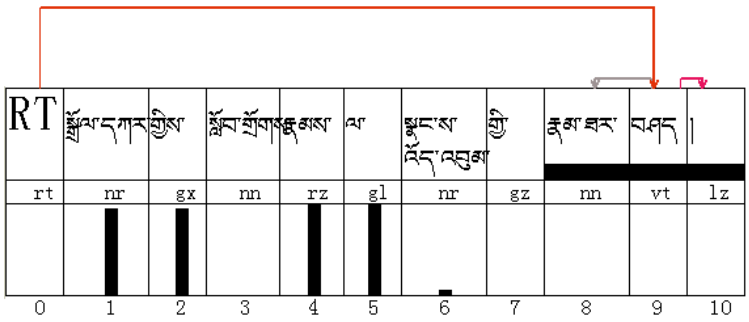


图 3 基于词对分类的半自动藏语依存句法标注图

3.2.1 词对依存边自动辅助提示

这部分工作是在确定词对依存分类的基础上进行的，主要完成词对依存边的标注(弧上关系标注)提供自动辅助提示功能，类似于词性标注。根据改进式子(4)得到词对分类依存边标注模型，$C_l(i,j)$ 表示依存词对 i 和 j 之间满足第 l 个依存边的概率值。$f(i,j,l)$ 是依存词对 i 和 j 之间特征向量，表示该词对间存在的边类型 l，其中 $l \in \{e-types\}$，*e-types* 表示所有可能的依存边，即依存关系类型集。如果一个特征 $f_k(i,j,l) \in f(i,j,e-types)$ 等于 1 表示该特征在训练语料中抽到的特征集中存在，否则不存在。 则词对分类依存边标注模型可定义为：

$$C_l(i,j) = \frac{\exp(w \cdot f(i,j,l))}{\sum_{e-types} \exp(w \cdot f(i,j,e-types))} = \frac{\exp(\sum_k w_k \cdot f_k(i,j,l))}{\sum_{e-types} \exp(\sum_k w_k \cdot f_k(i,j,e-types))} \quad (5)$$

依存边自动辅助提示中，本文依据表 1 给出的藏语依存句法规范，通过人工标注含有依存边信息 700 句句法树，利用表 2 给出的特征模板，包含最后一项依存边标注特征，用最大熵方法训练模型，使用训练出的最大熵模型进行词对弧上关系自动辅助提示和标注。见图 4 所示。当前选择的支配词是“རྟེན་ཐད་”，被支配词是“བཤད”。自动辅助提示的最佳依存边为直接宾语 OBJ。

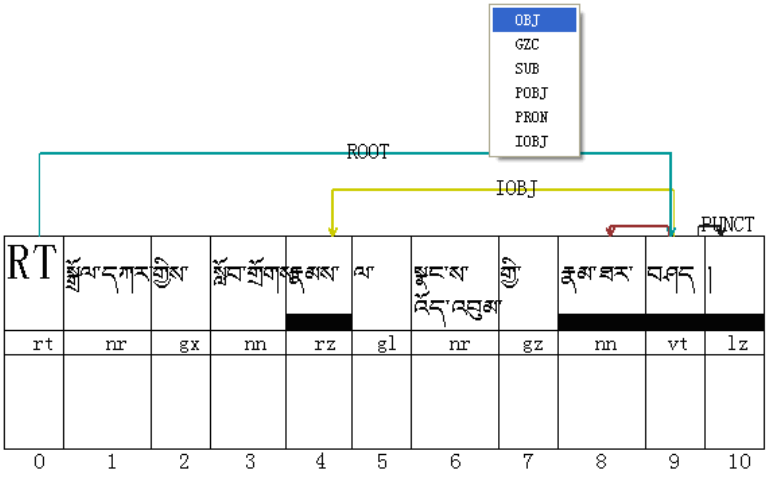


图 4 词对依存分类的边标注半自动辅助提示图

4. 树库构建及实验效果

为训练词对依存分类模型和词对依存边分类模型，本文人工分别构建了 2 千多句依存句法树和 7 百句依存边标注句法树库，用最大熵训练了模型，并结合这两种词对依存分类模型实现了半自动依存标注和修改工具 TibetanDepBuilder V2.4。利用此工具重新校对了人工标注的藏语依存句法树库 TDTreebank 1.0，其规模为 1 万句，平均长度小于 17 个藏语词，总词数规模为 16.7 万。整个树库的校对修改率达 10%，发现之前人工标注的主要错误包括：(1)复杂句句型的标注错误，包括从句结构，比较长的句子；(2)存在部分标注不一致。此外，本文利用半自动标注工具新标注了 1 千句句型比较复杂，且词数达 20 至 40 间的句法树。已加入到 TDTreebank 1.0 中，目前 TDTreebank 1.1 版本的规模为 1.1 万句。

表 3 校对树库语料

版本说明	训练语料（句）	测试集(句)	修改说明
TDTreebank 1.0	10000	300	校对前
TDTreebank 1.1	11000	300	校对并补充后

为了客观评价本文提出的词对依存分类半自动树库构建方法的效率，采用之前研发的判别式藏语句法分析为基线系统，重新校对并增加后的藏语依存树库作为训练语料。用之前的测试语料 300 句为测试集。以依存关系正确率(*depP*)、中心词正确率(*headP*)和整句完全依存正确率(*allP*)为性能分析指标[1]，对系统的藏语依存分析结果进行评价，给出了树库校对前后的评价指标，表 4 校对前后各项评测指标对比中正确率 I 为校对之前的评价指标，正确率 II 为校对后的各项评价指标。效果见图 5 所示。

表 4 校对前后各项评测指标对比

评测项	I 校对前正确率(%)	II 校对后正确率(%)
depP	81.20	84.26
headP	87.49	89.32
allP	34.58	35.87

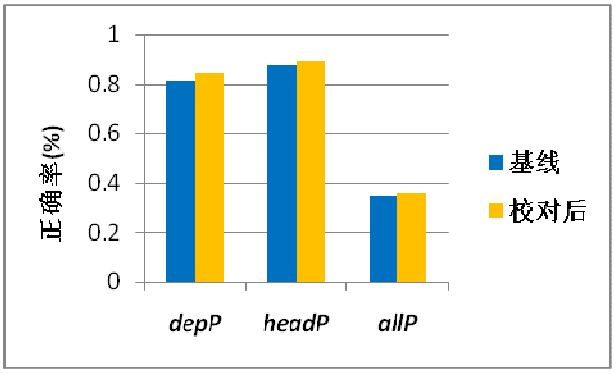


图 5　树库校对前后系统对测试语料的各项评价指标

5. 结论

针对藏语依存句法树库构建过程中存在的问题，本文提出词对依存分类的半自动依存句法树构建方法，描述了用于藏语依存句法结构及其标注规范，设计了词对依存分类模型和词对依存边分类模型，结合特征模板，分别在 2 千多句依存句法树和 7 百句依存边标注句法树上用最大熵训练了模型，用自顶向下标注模式实现了词对依存关系自动辅助提示和依存边类型自动辅助提示功能。

利用本文实现的词对依存分类半自动依存标注工具，校对了藏语依存树库 TDTreebank 1.0 后，经在同样测试集上实验显示，依存句法评测各项指标均有明显的提高。在很大程度上方便了句法分析树的校对，同时加快了藏语依存句法树库构建的进展。

参考文献

[1] 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.

[2] 胡书津. 简明藏文文法[M]. 昆明: 云南民族出版社, 1988.

[3] Peter Hellwig. Dependency Unification Grammar[C]. Proceeding of Coling'86. 1986.

[4] Marie-Catherine de Marne de and Christopher D. Manning. Stanford typed dependencies manual. 2008.

[5] 周明, 黄昌宁. 面向语料库标注的汉语依存体系的探讨[J]. 中文信息学报, 1994, 8(3): 35-51.

[6] 格桑居冕. 实用藏文文法[M]. 成都: 四川民族出版社, 1987.

[7] 华却才让,赵海兴.现代藏语依存句法标注初探[C].第十二届全国少数民族语言文字信息处理学术研讨会,2011.7.

[8] 才让加. 藏语语料库词语分类体系及标记集研究[J]. 中文信息学报, 2009, 23(4): 146-148.

[9] Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of COLING*, pages 340–345.

[10] Jiang Wenbin, Liu Qun. Dependency Parsing and Projection Based on Word Pair Classification[C]//Proc. of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: [s. n.], 2010: 12-20.

[11] McDonald R, Crammer K, Pereira F. Online Large-margin Training of Dependency Parsers[C]//Proc. of the 43rd Annual Meeting on Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2005: 91-98.

[12] Collins M. A New Statistical Parser Based on Bigram Lexical Dependencies[C]//Proc. of the 34th Annual Meeting on Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 1996: 184-191.

通讯作者：华却才让
联系电话：18911260587
联系地址：青海省西宁市城西区五四西路 38 号青海师范大学南院计算机学院
Email: cairanghuaque@aliyun.com