文章编号：1003-0077（2011）00-0000-00

基于多策略的藏语语义角色标注研究*

龙从军[1]，康才畯[1]，李琳[2]，江荻[1]

(1.中国社科院民族所，北京市 100081; 2.青海师范大学计算机学院，青海省西宁市 810004）

摘要：语义角色标注研究对自然语言处理具有十分重要的意义。英汉语语义角色标注研究已经获得了很多成果。然而藏语语义角色标注研究不管是资源建设，还是语义角色标注的技术探讨都鲜有报道。藏语具有比较丰富的句法标记，它们把一个句子天然地分割成功能不同的语义组块，而这些语义组块与语义角色之间存在一定的对应关系。根据这个特点，本文提出规则和统计相结合的、基于语义组块的语义角色标注策略。为了实现语义角色标注，文中首先对藏语语义角色进行分类，得到语义角色标注的分类体系；然后讨论标注规则的获得情况，包括手工编制初始规则集和采用错误驱动学习方法获得扩充规则集；统计技术上，选用了条件随机场模型，并添加了有效的语言特征，最终语义角色标注的结果准确率、召回率和 F 值分别达到 82.78%、85.71%和 83.91%。

关键词：藏语；语义角色标注；TBL；CRFs

中图分类号：TP391　　**文献标识码：**A

Multi-Strategic Research on Semantic Role Labeling of Tibetan

LONG Congjun[1], KANG Caijun[1], Li Lin[2], JIANG Di[1]

（Institute of Ethnology & Anthropology Chinese Academy of Social Sciences, Beijing 100081, China; The Computer College of Qinghai Normal University, Xining, Qinghai 810004, China）

Abstract: To study Semantic role labeling is of great significance for natural language processing. The researches of semantic role labeling about English and Chinese have obtained many achievements. However, the resources construction and technological means of semantic role labeling in Tibetan are still in initial stage. Tibetan has rich syntactic markers which naturally segment a sentence to different semantic chunking, and there are certain relationship between these semantic chunking and semantic roles. According to this characteristic, the authors of this paper propose the semantic role labeling strategy based on semantic chunking by combining two means of rules and statistics. In order to realize the semantic role labeling, the authors design classification system of Tibetan semantic roles and then discuss the acquisition of rules, including a manual initial rule sets and expanded rule sets from Transformation-Based Error-driven Learning (TBL). The authors adopt Conditional Random Fields (CRFs) Model. In processing of experiment, some effective language characteristics are added, and the results of semantic role labeling are 83.91% for precision, 82.78% for call rate and 85.71% for F values.

Key words: Tibetan; Semantic Role Labeling; TBL; CRFs

1 引言

自动语义角色标注（Semantic role labeling，缩写为 SRL）是自然语言处理的重要任务，对提高语言信息处理系统的性能具有重要的意义。语义角色标注的过程可以表述为：设立一套标签体系（角色分类体系），部分地标注句子的成分结构（能承载语义角色），使计算机自动的获得一定的“理解”能力。

最早研究 SRL 的是 Gildea 和 Jurafsky，他们开发了一套 SRL 系统，经对不同的两套语料测试，实验结果准确率分别约为 82%和 65%[1]。在 CoNLL2004 会议中，他们提交的论文强调对句法组块进行分类，在训练语料相同的情况下，比较了词到词与短语到短语的标注结果，

* **收稿日期：**　　**定稿日期：**

基金项目：国家自然科学基金重点项目(61132009)

测试结果为准确率 72.43%、召回率 66.77%和 F 值为 69.49[2]。CoNLL2007 会议，设立一个 Session 讨论 SRL[3]。CoNLL2008 又把 SRL 作为主要的评测任务，并以依存句法分析为基础，考察 SRL 和句法分析结果[4]。我国学者也较早地参与 SRL 研究，在 2005 的 CoNLL 会议的评测会上，刘挺等采用了最大熵模型进行英语 SRL 研究，以句法成分为标注单元，并加入了规则后处理模块，标注的结果为：准确率 79.65%，召回率 71.34%和 F 值 75.27[5]。汉语语义角色标注研究也取得了一系列的成果[6-8]，尤其是丁伟伟等人尝试了基于语义组块的汉语语义角色标注研究，获得了不错的效果[9]。

藏语目前没有公开可用的句法树库资源，难以采用句法成分和依存句法的语义角色标注研究。但是藏语具有比较丰富的句法标记，这些标记可以把一个藏语句子天然地切分成一定数目的块（语义组块），文献[10-13]从不同的角度对藏语的组块进行了讨论，但是未讨论组块与语义角色之间的对应关系。本文尝试采用统计和规则相结合的多策略的方法进行基于语义块的语义角色标注研究。规则的获得主要通过手工编制初始规则集，利用错误驱动学习的方法获得扩充规则集，然后利用这些规则对基于条件随机场模型的语义角色标注结果进行校正。

2 藏语的语义角色分类

2．1 藏语的标记

语义组块的分类需要关注两个问题：块的边界和块的类型，藏语中的标记兼顾了块边界识别与块的类型识别两大功能，下面以例句 1 来说明。

例 1:[ཁོ/rh][**ས་/wa**][ང/rhའི་/wgནང་ས་འཁོར་/ngགསར་ད/a][**ར་/wd**][བསྟོད་པ་ཐྱེད་/voཀྱིན་ཡོད་/t/xp]

khos ngavi rlangs vkhor gsar bar bstod pa byed kyin yod.

他对我的新汽车夸赞。

句子中下划线部分分别是施格和对象格标记，两个标记把这个句子划分成三个块；同时ས་和 ར་也表明了它们前面组块的角色类型，一个是施事，一个是对象。但是标记具有多功能性，同样一个 ར་可能表示对象，也可能表示处所。这给语义角色类型的确定带来一定的困难，本文在词性标注语料基础上开展语义角色标注研究，不同语境下的标记功能是确定的，这避免了标记的歧义问题。标记分为两个大类，格标记和助词标记，格标记包括施事格、工具格、时间格、处所格、对象格、从格、领有格、比较格、向格、使动格、伴随格等，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。

2．2 藏语语义角色分类原则和体系

SRL 具有语言工程研究特点，语义角色的分类将应用到语料标注中去，实践性强。如果分类过细，对语言研究来说当然是件好事，但在实际语料标注中，会遇到很多困难。如果分类过粗，也许达不到研究的目的。因此一个大致合适的语义角色分类体系是十分必要的。袁毓林专门对语义角色分类的精细等级进行了探讨，从现有的语义角色分类研究的结果可以粗略的分为微观、中观和宏观三个等级。微观分类包括基于特定动词和基于特定领域的语义角色。中观层级的分类包括各种语义格的分类，它是基于动词类而不是具体的动词。宏观层级的分类包括原型施事(Proto-Agent)与原型受事(Proto-Patient) [14]。藏语的语义角色分类需要关注格及助词的标记特征以及动词的类型特征，因此分类的结果与中观层级的分类相近。同时本文作者还拟定了如下的三个分类原则：

（1）参照性原则。所谓参照性就是充分借鉴现有英、汉语中语义角色分类体系的成功经验[15-20]，结合藏语语言实际特点，加工改造。

（2）注重标记特征。一种语言中的句法标记特征越多，计算机对这种语言的句法理解就更容易把握。藏语与汉语相比，以格标记为核心的句法标记十分丰富，是语义角色分类的重要依据。表 1 列示了句法标记与可能的语义角色的对应关系。

表 1 句法标记与语义角色对应关系

标记	语义角色	标记	语义角色
-ས/གིས/ཀྱིས/ཀྱིས/ཡིས	施事(AG)	ངོ	涉事(CE)
ལ་/-ར	领事(GE)	ལ/-ར/སུ/ར/ཏུ	结果(OE)
-ས/གིས/ཀྱིས/ཀྱིས/ཡིས	使役(CA)	-ས/གིས/ཀྱིས/ཀྱིས/ཡིས/ནས	方式(MR)
ལ/-ར/སུ/ར/ཏུ	受事(PT)	-ས/གིས/ཀྱིས/ཀྱིས/ཡིས	工具(IT)
ལ/-ར/སུ/ར/ཏུ	受役(BE)	-ས/གིས/ཀྱིས/ཀྱིས/ཡིས	材料(ML)
ནས/ལས	起源(SE)	ལ/-ར/སུ/ར/ཏུ	时间(TM)
ལ/-ར/སུ/ར/ཏུ	对象(TT)	ལ/-ར/སུ/ར/ཏུ	处所(LC)
ནས/ལས/བས	依据(BS)	ལ/-ར/སུ/ར/ཏུ	方位(DN)

（3）角色配对原则。句法标记并不能与语义角色构成一一对应关系，仍然存在部分无标记的语义组块。针对这种情况，本文作者在对语义角色分类时充分考虑了无标记与有标记的语义角色块在一个句子中的配对关系，如施事与受事，领事与属事，系事与类事，使役与受役之间存在配对关系；特殊句型与语义角色的关系，如领事、属事与领有句有关，系事与类事与判断句有关，使役与受役与使动句有关等。

综合以上的各种因素，本文作者最终为藏语设计了 22 个语义角色类型，具体如表 2 所示。

表 2 藏语语义角色分类体系

名称	标记	名称	标记	名称	标记	名称	标记
施事	AG	属事	BL	起源	SE	方式	MR
受事	PT	系事	CP	对象	TT	工具	IT
当事	EX	类事	IS	依据	BS	材料	ML
客事	BO	使役	CA	涉事	CE	时间	TM
领事	GE	受役	BE	结果	OE	方位	DN
处所	LC	目的	PU				

3 语义角色标注规则构建

与统计方法相比，规则方法在自然语言处理中并无明显优势，这对于资源丰富、数据获取便利的大语种来说更是如此。但是对于资源少、句法标记较丰富的藏语来说，在现阶段也不失为一种有益的尝试。为此，本文采用手工编制初始规则集和利用基于转换的错误驱动学习算法(Transformation-Based Error-driven Learning, TBL)对规则库进行泛化，从而获得扩充规则集。

3.1 初始规则集

初始规则包括语义块边界规则和语义角色与格标记及助词的对应规则。规则的获得主要由人工总结归纳。语义块边界规则由左边界、右边界、双边界和左右边界例外四个部分组成。其中左右边界例外是一个调节规则，就是对左右及双边界标注结果进行纠错。四个部分规则共有 271 个，右边界特征 114 个，双边界特征 119 个，左边界特征 15 个（包括全部动词和否定副词等特征词），左右边界例外特征 35 个。语义角色与格标记及助词的对应规则 63 个。

3.2 扩充规则集

在初始规则集的基础上，本文作者采用 TBL 算法自动从语料中学习并建立扩充规则集。TBL 算法利用学习器从语料中自动获取转换规则集，因此建立一个高效的学习器是 TBL 算法的关键。学习所需资源主要包括以下三方面：（1）正确标注语义角色的语料，（2）经初始标注的语义角色语料，（3）转换规则模板集合。通过比较资源（1）和资源（2）之间的标注差异，习得扩充规则集。

4 统计模型及特征选择

4.1 条件随机场模型

条件随机场（Conditional Random Fields，CRFs）是一种判别式概率模型，多用于标注或者分析序列材料。在基于统计的标注方法中，条件随机场模型具有很好的效果，其模型思想主要来源于最大熵模型，但又不存在最大熵模型的数据稀疏问题；同时也无需对数据进行不必要的独立性假设，在这个方面也优于隐马尔科夫模型（HMM）。CRFs 通常采用如图 1 的一阶链式结构。

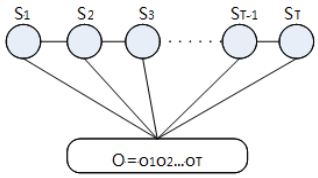

图 1 CRFs 链式图

4.2 标签设计

语义角色标注的标签设计十分关键，在训练语料不丰富的前提下，标签的数量会直接影响标注的效果。本文的语义角色标注所使用的语料经人工标注，标注语料能够提供词性信息、语义组块边界信息和语义角色标记信息。如果把这些信息结合起来考虑，设计一套联合标签，在一定程度上，可能会提高标注效果。边界标记使用了 BIO 标记法，单词成块或者块外标点为 S，块开始用 B，块结尾用 E，块中间用 I。

表 3 语义角色标注联合标签

词形	ཁོ་	མ་	ར་	ལ་	རྒྱངས་འགོངས་	གནས་ཚུལ་	ར་	བསྒྲུབ་པའྱུང་	ཡིན་ཡོད་	།
词性	rh	wa	rh	wg	ng	a	wd	v	t	xp
边界标记	S	S	B	I	I	E	S	B	E	S
角色标记	AG	M	TT	TT	TT	TT	M	P	P	OT

语义角色以角色类型作为标签。联合标签还包括词形和词性，具体情况如表 3 所示。

4.3 特征选择

条件随机场算法中，数据的很多重要信息是通过特征函数给出的。CRF 工具以特征模板的方式，给出这些特征函数的定义，使选取特征和定义特征函数非常方便。本文所使用的特征包括词形信息、词性信息、语义组块音节长度信息、谓语动词类型信息以及谓语动词与语义组块之间的距离信息。这些信息的具体描述如下：

词形：指词的形式属性，如 ཁང（khong,他）与 ཁྱེད་རང（khyed rang,您）是两个不同的词形。

词性：指词的词类信息；ཁང（khong,他）、ཁྱེད་རང（khyed rang,您）是人称代词（rh），而 གནས་ཚུལ（gnas tshul,情况）是普通名词（ng）。

音节长度：是指一个语义组块的音节数量，语义组块的平均音节数量可以作为语义组块边界识别的参考。

[གནས་ཚུལ/ngམི་འདྲ་བ/iaའི་/wgབོག་/nd] [ཀོ་བ་/ngམི་འདྲ་བ་/ia][ཡོད/ve][།/xp]（gnas tshul mi vdra bavi vog go ba mi vdra ba yod, 不同的情况下有不同的理解。）

这个句子中有四个块，其中[ཡོད/ve]（yod,有）是谓语组块，[།/xp]是标点符号，不需要考虑，其余两个语义组块各自的音节长度分别是 7 个和 4 个。

谓语动词类型：是指谓语动词的语义类型，本项研究按照动词必有论元的数量对动词分类，分为一元动词、二元动词和三元动词。动词语义类型可以影响动词携带语义角色的数量。

谓语动词与语义块的距离：是指谓语动词与承载语义角色块之间间隔的音节数，一般来说受事语义角色与谓语动词近，施事语义角色离动词远，这些特征可能有效地辅助推断语

义角色。

5 实验及结果分析

5.1 工具与语料

本文使用由日本松本实验室的 Taku Kudo 博士开发的 CRF++软件包[21]。首先构建一个 Baseline 系统，该系统只选择词形及词性作为特征，其识别结果作为后续实验比较的基准。在后续实验中增加新特征，通过与 Baseline 结果的比较，确定添加的特征对语义块的识别是否有效。添加的特征分别是语义块音节长度、谓语动词类型和谓语动词与语义块的距离三个特征。选取约 5000 个手工标注语义组块边界和语义角色类型的句子作为训练语料，用同类型的 500 个句子作为测试语料来检验算法的性能。

5.2 实验结果评价及结果分析

本文采用语言信息处理通用评价指标来检验语义组块标注结果的效率，这些指标是准确率（P）、召回率（R）和 F 值，计算公式如下。

$$P = \frac{\text{正确语义组块数}}{\text{召回组块总数}} \times 100\%$$

$$R = \frac{\text{正确语义组块数}}{\text{语义组块总数}} \times 100\%$$

$$F = \frac{2 \times P \times R}{P + R} \times 100\%$$

实验结果表明，利用词形和词性的 baseline 模板的语义角色标注准确率、召回率和 F 值分别为 68.88%、63.10%和 64.85%。当分别加入语义组块音节特征、谓语动词类型特征和语义块与谓语动词的距离特征后，获得的标注效果都有明显的提升，其中加入语义组块音节特征效果最为明显，准确率、召回率和 F 值分别提高到 78.60%、71.34%和 74.80%。第二个具有明显作用的特征是谓语动词与语义角色块的距离特征，谓语动词类型特征同样也能提高标注效果，具体情况如图 2 所示。

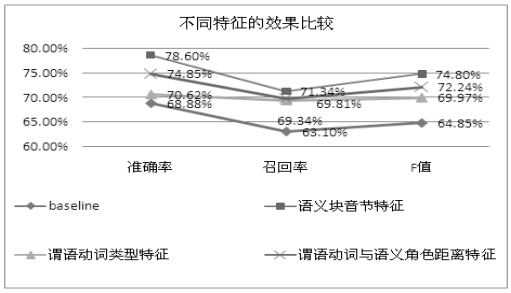

图 2 加入不同特征测试指标的变化情况

基于统计的方法在边界识别与语义角色标注中出现的错误类型可以概括为如下几种类型：

（1）边界识别错误。如例句 2：

例 2：[ཀུང་ཁྲི་/ng]{BS}[ལྟར་/ua][ན་/c][དེ་/rd]{CP}[ཀ་ཚོད་/rw]{IS}[རེད་/vl][དམ/y][།/xp]（kung khri ltar na de ga tshod red dam. 按照公尺算的话，这些是多少？）（参考答案）

这个句子的边界识别结果为：

[ཀུང་ཁྲི་/ngལྟར་/ua][ན་/c][དེ་/rd]{CP}[ཀ་ཚོད་/rw]{IS}[རེད་/vl][དམ/y][།/xp]（语义角色标注结果）

其中 ལྟར་/ua 是一个边界标记词，边界标记错误，导致{BS}语义角色标注失败。

例 3：[ཁྱེད་རང་/rhགི་/wgངའི་/ng]{PT}[ང་/rh]{TT}[ཨཱ་/wd][ཞོད་/vo][དང་/y][།/xp]（khyed rang gi ming nga la shod dang. 对我说你的名字。） （参考答案）

这个句子的边界识别结果为：

[ཁྱེད་རང་/rhགི་/wgངའི་/ngང་/rh]{TT}[ཨཱ་/wd][ཞོད་/vo][དང་/y][།/xp] （语义角色标注结果）

语义组块[ཁྱེད་རང་/rhགི་/wgམིང་/ng]（khyed rang gi ming, 你的名字）和[ང་/rh]（nga, 我）边界识别错误，导致{PT}语义角色标注失败。

（2）边界识别正确，语义角色标注错误。语义角色标注的错误表现有：语义角色未标注，语义角色标注的位置错误和语义角色选择错误。如例句 4、5。

例 4: [གནས་ཚུལ་/ngཚང་མ་/a]{PT}[ཁོང་/rh]{TT}[ལ་/wd][ཤོད་/vo][དང་/y][།/xp] （gnas tshul tshang ma khong la shod dang. 全部的情况对他说吧。） （参考答案）

[གནས་ཚུལ་/ngཚང་མ་/a][ཁོང་/rh]{TT}[ལ་/wd][ཤོད་/vnདང་/y][།/xp] （语义角色标注结果）

组块[གནས་ཚུལ་/ngཚང་མ་/a]（gnas tshol tshang ma, 全部情况）边界识别正确，但是{PT}语义角色标注失败，属于语义角色未标注。

例 5: [ཉིན་མོ་/nt]{EX}[རིམ་བཞིན་/d][ཐུང་/a]{OE}[དུ་/ub][འགྲོ་/vo][གི/t][།/xp]（nyin mo rim bzhin thung du vgro gi. 白天逐渐变短了。） （参考答案）

[ཉིན་མོ་/nt][རིམ་བཞིན་/d]{EX}[ཐུང་/a]{OE}[དུ་/ub][འགྲོ་/vo][གི/t][།/xp] （语义角色标注结果）

语义组块[ཉིན་མོ་/nt]与[རིམ་བཞིན་/d]边界识别正确，但是{AG}语义角色标注错误，其中[རིམ་བཞིན་/d]不能承载语义角色，属于非语义角色组块，本例属于语义角色标注位置错误。

5. 3 实验结果改进

利用已经建立的边界规则库和语义角色标注规则来优化统计标注的结果，在训练语料规模有限的情况下可能会产生一定的效果。因此我们利用手工编制初始规则集和利用 TBL 方法获得的扩充规则集对统计结果进行校正。在进行第二次实验时，本文选择了统计标注最好结果作为规则校正的对象。图 3 表示 baseline 统计方法标注、加入语义组块音节数统计方法标注以及统计和规则相结合的标注三种实验结果的对比。

从图 3 可以看到，利用规则方法对统计结果进行校正，与单纯依靠统计方法相比，实验结果有大幅度地提升，准确率、召回率和 F 值分别达到了 82.78%、85.71%和 83.91%，可见规则方法的调节效果还是比较明显的。

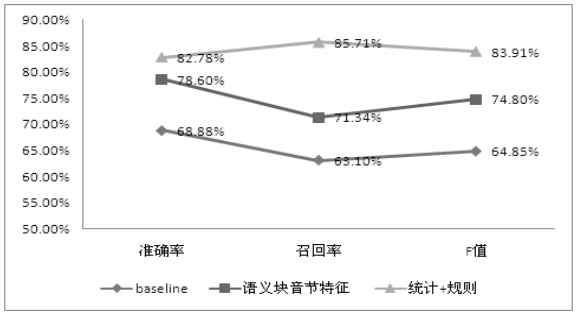

图 3 三种方法标注结果对比图

6 结束

语义角色标注研究属于浅层句法分析的主要内容之一，在当前完全句法分析存在诸多困难的情况下，开展浅层句法分析可以有效地提高机器分析与理解自然语言的能力。语义角色标注研究的成果在机器翻译、自动问答、信息抽取等诸多领域都可以得到广泛使用。本文探讨了藏语语义角色标注研究，通过利用统计和规则相融合的策略，提升了语义角色标注的效果，实验结果准确率达到了 82.78%。但是本项研究中，对于嵌套语义组块和长距离语义组块的标注效果并不理想，这类错误拉低了标注的准确率。在后续研究中，除了扩充大规模的训练语料和精细化规则集之外，还需要对嵌套语义组块和长距离语义组块进行专门的纠错处理。

参考文献

[1] Daniel Gildea, Daniel Jurafsky: Automatic Labeling of Semantic Roles, Computational Linguistics, Volume 28, Number 3,p1-45.

[2] Kadri Hacioglu, Sameer Pradhan, Wayne Ward, James H. Martin and Daniel Jurafsky, Semantic Role Labeling by Tagging Syntactic Chunks. In Proceedings of ConNLL-2004.
[3] http://www.cs.jhu.edu/EMNLP-CoNLL-2007/.
[4] http://www.clips.ua.ac.be/conll2008/.
[5]Ting Liu, Wanxiang Che, Sheng Li, Yuxuan Hu and Huaijun Liu,Semantic Role Lableing System using Maximum Entropy Classifer, In Proceedings of ConNLL-2005.
[6]Yu Jiangde, Fan Xiaozhong ect. Semantic role labeling based on conditional random fields, Journal of Southeast University. Vol 23, No. 3, p 361- 364.
[7]王步康、王红玲等：基于依存句法分析的中文语义角色标注[J]，中文信息学报，2010 年 1 月.
[8]刘挺、车万翔等：基于最大熵分类器的语义角色标注[J], 软件学报，2007 年 3 月，Vol18, No.3.
[9]丁伟伟、常宝宝：基于语义组块分析的汉语语义角色标注，中文信息学报，2009 年 9 月.
[10]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:p160-166.
[11]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[A]， 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，p13-106.
[12]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 年第 6 期.
[13]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.
[14]袁毓林：语义角色的精细等级及其在信息处理中的应用[J]，中文信息学报，2007 年 7 月，P10-20.
[15]周强、詹卫东、任海波：构建大规模的汉语语块库[A]，清华大学出版社：自然语言理解与机器翻译，2001，pp102-107.
[16]周强、孙茂松：汉语句子的组块分析体系[J]，计算机学报，1999, 22(11), pp1158-1165.
[17]杨敏、常宝宝：基于北京大学中文网库的语义角色分类[J]，中文信息学报 2011 年 3 月， vol.25，No.2.
[18] http://www.keenage.com/.
[19]鲁川：汉语语法的意合网络，商务印书馆，2001 年，P111.
[20]林杏光：词汇语义和计算语言学，语文出版社，1999 年 5 月，P184.
[21]http://CRFspp.googlecode.com/svn/trunk/doc/index.html.

作者简介：

龙从军（1978——），男，博士，主要研究领域：藏语语法、藏语信息处理。Email:longcj@cass.org.cn.
康才畯（1980——）男，博士，主要研究领域：藏语信息处理。
李琳（1980——），女，博士，主要研究领域：藏语信息处理。
江荻（1954——）男，研究员，主要研究领域：藏语语法、藏语信息处理。