

藏语语义角色研究综述

赛毛吉

西北民族大学中国民族信息技术研究院, 甘肃 兰州

收稿日期: 2025年4月25日; 录用日期: 2025年6月6日; 发布日期: 2025年6月19日

摘要

藏语作为汉藏语系的重要语言, 具有丰富的形态变化和独特的语法结构, 其语义角色研究对于藏语自然语言处理、机器翻译、信息抽取等领域具有重要意义。本文系统梳理了藏语语义角色研究的发展历程、理论框架、方法创新以及应用前景, 旨在为相关领域的研究者提供全面的参考。文章首先回顾了藏语语义角色研究的理论基础, 包括传统文法与现代语言学理论的结合; 然后详细分析了藏语语义角色体系的构建与分类标准; 接着总结了基于规则和统计学习的标注方法及其性能表现; 最后探讨了藏语语义角色研究面临的挑战与未来发展方向。通过对现有文献的系统分析, 本文揭示了藏语语义角色研究从理论探索到实际应用的完整发展路径, 为后续研究提供了方法论指导。

关键词

语义角色, 藏语, 自然语言处理, 语义角色标注

A Review of Semantic Role Research in Tibetan

Maoji Sai

China Academy of Information Technology for Nationalities, Northwest Minzu University, Lanzhou Gansu

Received: Apr. 25th, 2025; accepted: Jun. 6th, 2025; published: Jun. 19th, 2025

Abstract

As an important language in the SinoTibetan language family, Tibetan features rich morphological changes and unique grammatical structures. The study of semantic roles in Tibetan is of great significance for fields such as natural language processing, machine translation, and information extraction of Tibetan. This paper systematically reviews the development history, theoretical framework, methodological innovations, and application prospects of semantic role research in Tibetan, aiming to provide a comprehensive reference for researchers in related fields. The article first

reviews the theoretical basis of semantic role research in Tibetan, including the integration of traditional grammar and modern linguistic theories; then it analyzes in detail the construction and classification standards of the semantic role system in Tibetan; next, it summarizes the annotation methods based on rules and statistical learning and their performance; finally, it discusses the challenges and future development directions of semantic role research in Tibetan. Through a systematic analysis of existing literature, this paper reveals the complete development path of semantic role research in Tibetan from theoretical exploration to practical application, providing methodological guidance for subsequent research.

Keywords

Semantic Role, Tibetan, Natural Language Processing, Semantic Role Labeling

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

语义角色标注(Semantic Role Labeling, SRL)作为自然语言处理领域的基石任务,致力于挖掘谓词与论元之间深层次的语义关联,是构建语义理解系统的核心环节。藏语作为汉藏语系中极具代表性的语言,承载着藏族千年的历史文脉,其丰富的形态变化与独特的语法结构,为语义角色研究提供了独一无二的视角。深入开展藏语语义角色研究,不仅有助于推动藏语自然语言处理技术的发展,对于理解汉藏语系语言的共性与差异,乃至促进多语言信息处理技术的整体进步,都具有深远意义。

藏语的语言特性使其语义角色研究具备独特价值。在语法层面,藏语拥有一套完备且复杂的格标记系统,语义角色常通过特定的格助词精准标示,例如施事通常由“གཞི”表示,受事则多由“ལ”体现。这种显性的语法标记,在一定程度上为语义角色的识别提供了明确线索,但也因藏语动词的多样性而增加了研究难度。藏语动词依据及物性、自主性、使动性等多个维度可划分成不同类别,这些类别差异直接影响论元结构的配置,使得同一语义角色在不同动词框架下可能呈现出不同的表现形式。

回顾藏语语义角色研究历程,其经历了从理论探索到实践应用的蜕变。早期研究受国际语义角色标注体系的启发,学者们尝试将英语 PropBank 和汉语 CPB 等标注体系应用于藏语研究,为后续发展奠定了基础。随着研究的深入,越来越多的学者意识到藏语自身特点的重要性,开始致力于构建符合藏语特色的语义角色体系,如完么扎西等[1]提出的 21 种角色分类体系,推动藏语语义角色研究迈向新的阶段。在技术方法上,也实现了从传统基于规则方法向先进深度学习模型的跨越。

2. 理论基础与体系构建

2.1. 理论渊源

藏语语义角色研究的理论根基,深深扎根于传统藏文文法与现代语言学理论的交融之中。藏文传统文法经典著作,如被誉为“藏语文法之母”的《三十颂》,对藏语句法成分、格标记系统以及语言结构规律进行了系统且深入的描述,为现代语义角色研究提供了丰富的语言素材与理论雏形。而现代语言学中的论元结构理论和框架语义学,则为藏语语义角色研究提供了科学的分析工具和理论框架,使研究者能够从全新的视角剖析藏语的语义结构。

藏语动词分类理论在语义角色研究中占据核心地位。藏语动词根据及物性、自主性和使动性等维度

进行细致分类，不同类型的动词对论元结构的要求各不相同。例如，自主及物动词在句子中通常需要施事和受事两个核心论元来完整表达语义，如“**ལུག་གིས་རྩ་ཚོས་** lug gis rtswa zos (羊吃草)中，“**ལུག་**”(羊)为施事，“**རྩ་**”(草)为受事；而不及物动词往往只需要一个论元，如“**མིག་གི་མཐོང་** mig gi mthong (眼睛看见)，“**མིག་**”(眼睛)作为唯一论元，承担着动作执行者的角色。

2.2. 分类体系演进

藏语语义角色分类体系的构建经历了从简单到复杂、从通用到专用的发展过程。早期研究倾向于直接采用或轻微调整英语或汉语的语义角色体系。见表 1，珠杰等[2]在 PropBank 基础上研究藏语特殊语义角色标记规范和标记方式，保留了 PropBank 的核心角色如 Agent (施事)、Patient (受事)等，同时针对藏语特点增加了一些特殊角色标记。随着研究的深入，学者们开始构建更加符合藏语特点的语义角色体系。祁坤钰[3]采用论元结构分析理论，以藏语逻辑格为对象，详细分析了受动格、施动格、目的格、来源格和处所格所包含的语义角色，提出了一个藏语语义角色体系的雏形。这一研究的重要意义在于，它不再简单套用其他语言的分类标准，而是从藏语自身的格标记系统出发，探索语义角色的内在规律。最新研究成果展示了藏语语义角色分类体系的精细化趋势。完么扎西[1]等提出了包含 21 种角色的藏语语义角色分类体系，该体系根据动词的及物性、自主性和使动性等范畴，对藏语谓词论元结构进行了详细探讨。还有旦正才让[4]的硕士研究，从标注语料中提取了 46 个具有藏语逻辑格语义指向的语义角色，并将其分为 19 个核心论元角色和 27 个准角色(修饰性成分)，然后将准角色合并到核心角色的相应属性类别中。这种核心边缘的二分法既反映了不同语义成分与动词关系的紧密程度，又保持了体系的简洁性和实用性。

Table 1. A classification system of Tibetan semantic roles proposed by different scholars
表 1. 不同学者提出的藏语语义角色分类体系

研究者	角色数量	分类依据	主要特点	代表性角色
珠杰等(2018)	未明确总数	PropBank 规范	国际通用框架本地化	Agent, Patient, Location
祁坤钰(2014)	雏形体系	藏语逻辑格	基于格标记系统	施动格角色、受动格角色
完么扎西等(2023)	21 种	动词语义类别	动词中心论	施事、受事、工具、受益者
旦正才让(2023)	46 种(19 核心)	语料统计分析	核心边缘区分	核心：施事、受事；准角色：时间、方式

2.3. 语义角色与句法实现

藏语语义角色与句法成分之间存在着紧密且相对明确的对应关系。从格标记角度来看，施动格标记“**ཀྱིས་**”通常用于表示施事角色，强调动作的发起者；例如，在“**ཨ་མེས་ཀྱིས་གནང་གཏམ་བཤད།**” a mes kyis gnav gnam bshad (爷爷讲故事)中，“**ཨ་མེས་**”(爷爷)是施事角色也是主语，“**ཀྱིས་**”标记为施事格，“**གནང་གཏམ་**”(故事)是涉事对象也是涉事宾语，“**བཤད།**”(讲)是动词也是谓词；受动格标记“**ལ་**”除了常表示受事角色外，在一些语境中还可表示对象角色。而在“**སྒྲུབ་མཁས་པ་མ་ལ་ལ་ཐ་མ་བསྐོས།**” sgröl mas pha ma la za ma bskol (卓玛给父母做饭)中，“**ལ་**”标记为格标记，“**ཐ་མ་**”表示对象角色。此外，动词自身的词汇语义对语义角色的句法实现方式有着重要影响，不同语义的动词可能要求论元以不同的句法形式呈现，进一步增加了藏语语义角色研究的复杂性。

3. 标注方法与技术实现

藏语语义角色研究从理论构建到实际应用的关键环节在于标注方法的创新与技术实现的有效性。本部分将系统梳理藏语语义角色标注的技术发展脉络，分析基于规则、统计学习以及深度学习的各种方法的特点与性能，并探讨藏语语言特性对这些技术方案选择的影响，从而揭示藏语语义角色自动标注的技

术演进路径。

3.1. 基于规则的方法

在藏语语义角色研究初期，基于规则的方法占据主导地位。以珠杰[2]等学者构建的规则系统为例，其构建过程主要包括以下步骤：首先，广泛收集涵盖各种句式和语义场景的例句，形成大规模的例句库，确保规则的覆盖范围；其次，通过人工方式对例句的句法结构和语义构成规律进行细致标注和分析，总结出一般性的规则；最后，依据这些规则制定语义角色标注的具体规范和流程。

龙从军等[5]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%。

3.2. 统计机器学习方法

随着研究的推进，统计机器学习方法逐渐应用于藏语语义角色标注领域，其中条件随机场(CRF)是应用最为广泛的方法之一。见表 2，在使用 CRF 进行标注时，常用的特征包括词形、词性、格助词、动词形态、位置信息以及各种组合特征等。

将配价信息融入 CRF 模型，能够显著提升系统性能。通过考察动词的配价数，可以准确判断核心论元的数量；而配价类型则会影响论元在句法结构中的实现位置。例如，二价动词通常需要两个核心论元，且不同配价类型的动词，其论元的排列顺序和句法表现形式可能存在差异。

Table 2. Tibetan semantic role annotation commonly used features

表 2. 藏语语义角色标注常用特征

特征类型	具体示例	语言学意义	适用模型
词形特征	当前词、前后词	词汇本身信息	CRF、BiLSTM
词性特征	名词、动词、格助词	语法范畴信息	CRF、BiLSTM
格标记	施动格“གཞི”、受动格“ལེ”	语义角色线索	规则、CRF
动词形态	及物性、自主性	论元结构预测	规则、CRF
位置特征	与动词的距离	句法关系指示	CRF、BiLSTM
组块信息	名词短语边界	语义单元识别	规则、CRF

3.3. 深度学习方法

近年来，深度学习方法在藏语语义角色标注中展现出强大的优势，其中 BiLSTMCRF 架构成为当前主流。BiLSTM 能够充分学习句子的上下文信息，从前后两个方向捕捉词语之间的语义依赖关系，有效解决了传统方法对上下文信息利用不足的问题；CRF 则通过对标注序列进行全局优化，确保标注结果符合语言的语法和语义规则，进一步提高标注的准确性。同时，该架构减少了对人工设计特征的依赖，能够自动从大量语料中学习到有有效的语义和语法信息。

旦正吉[6]针对藏语句子语义分析和理解的困难，以及语义种类繁多、歧义等难点，提出了一种基于藏语音节向量(Tibetan syllables, TS)的双向长短时记忆网络 II (Bidirectional Long Short-Term Memory, Bi-LSTM)和条件随机场(Conditional Random Field, CRF)相结合的深度学习混合模型，用于藏语语义块识别。

基于人工标注的藏语语义块构建了 TS-BiLSTM-CRF 模型, 利用 BiLSTM 作为特征提取器, 并使用 CRF 对语料集进行序列标注。为促进藏语语义块识别人工智能领域内的音节向量化表示的多元化, 提出了一种融合 BERT 藏语音节嵌入的双向 LSTM-CRF 模型。

4. 资源建设与实证研究

藏语语义角色研究从理论探索到实际应用的转化, 离不开高质量的标注资源建设和系统的实证研究。本部分将全面考察藏语语义角色标注语料库的建设现状与实践经验, 分析不同标注体系的性能表现与适用范围, 并探讨语义角色研究在藏语自然语言处理中的具体应用场景, 从而揭示资源建设与理论研究、技术应用之间的互动关系。

4.1. 语料库建设

在藏语语义角色研究的资源建设方面, 旦正才让[4]构建的标注语料库具有重要意义。该语料库以五省涉藏地区语文教材为主要来源, 涵盖了丰富的语言场景和内容。语料库规模达到 8600 个句子, 包含 136,605 个词, 采用结构化序列标注格式, 便于数据的存储、管理和分析。在角色体系方面, 该语料库采用 46 个角色(19 个核心角色 + 27 个准角色)的分类体系, 能够全面、细致地标注藏语语义角色。

通过对该语料库的统计分析发现, 藏语句子平均长度为 13 个词, 每个句子平均包含 2 个语义角色; 约 29.15% 的名词在句子中承担语义角色。这些统计特征为深入了解藏语语义角色的分布规律和语言特点提供了数据支持, 也为后续的研究和模型训练提供了重要参考。

4.2. 应用研究

藏语语义角色研究的成果在多个领域展现出应用潜力。在句法和语义联合标注方面, 通过语义角色标注能够为句法分析提供更丰富的语义信息, 使句法分析结果更加准确和合理; 在机器翻译领域, 语义角色信息有助于提升译文的质量, 使译文在语义表达上更加准确和自然; 在信息抽取与问答系统中, 语义角色标注能够帮助系统更准确地理解用户问题, 提取关键信息, 提高回答的准确性和相关性; 在语言教育与词典编纂方面, 语义角色研究成果可为教学内容设计和词典释义提供参考, 帮助学习者更好地理解 and 运用藏语。

4.2.1. 机器翻译中的语义角色应用

在汉藏机器翻译中, 语义角色可有效解决若干关键难题。藏汉机器翻译技术跟汉英机器翻译技术有所不同, 其中, 很重要的一个方面, 藏语更依赖于格助词等虚词在句子中的作用, 格助词种类繁多, 用法差异很大。何向真等[7]针对藏语格助词进行分析, 在藏语短语句法树库的基础上, 加入了藏语本体特征的语义信息, 形成融合藏语语义信息的藏汉机器翻译方法。通过对比基于短语和句法的实验分析, 该方法可以很好地应用于藏汉机器翻译系统。藏语通过格助词标记论元角色, 而汉语依赖语序和介词, 我们以“把”字句翻译为例进行深入分析:

例子: 例如藏语 “ངས་དཔེ་ཆ་ཁོ་ལ་བྱིན།” ngas dpe cha kho la byin .(我把书给了他) “ལ” 标示受事, 传统统计方法常错误翻译为“我他给了书”, 忽略了“把”的受事标记功能。基于 SRL 的方法首先识别出: 谓词: 给, 施事: 我(ངས་), 受事: 书(དཔེ་ཆ), 接受者: 他(ཁོ), 格标记: (ལ), 然后根据角色格标记对应关系生成正确译文: 我把书给了他。

4.2.2. 信息抽取方面的应用

万福成等[8]在藏语短语句法分析的基础上, 融入功能语义信息线索, 采用在藏语短语句法树库的基础上加入语义角色标注的方法, 并融合事件触发词信息, 对藏文信息进行抽取, 并通过实验进行了对比

分析。事件特征融合与已标注信息对藏文信息抽取有帮助，模型可以很好地应用于藏文信息抽取工作。

例如：从藏语新闻报道中自动提取事件要素，首先输入文本：

“ཟླ5ཚེས15ཉིན་རྒྱལ་ས་པེ་ཅིང་དུ་རྒྱལ་ཡོངས་མི་དམངས་འཐུས་ཚིགས་ཀྱི་འཐུས་མིས་ཐུག་ཁྱིམ་གྱིས་བསྐྱར་རྒྱལ་ཚོམ་བྱས།” zla 5 tshes 15 nyin rgyal sa pe cing du rgyal yongs mi dmangs vthus tshogs kyi vthus mis jus gzhir gros bsdur rags tsam byas. (5月15日在首都北京，全国人大代表初步讨论了方案)，然后在模型内部进行语义角色标注：谓词：讨论(ཐུག་ཁྱིམ་བྱས་)、施事：全国人大代表(རྒྱལ་ཡོངས་མི་དམངས་འཐུས་མིས་)、时间：5月15日(ཟླ5ཚེས15ཉིན་)、地点：北京(པེ་ཅིང་)、受事：方案(ཐུག་ཁྱིམ་བྱས་)，输出结构化结果：{“事件类型”：“会议讨论”，“时间”：“20230515”，“地点”：“拉萨”，“参与者”：“全国人大代表”，“讨论内容”：“方案”}。即使在多个方面藏语语义角色标注(SRL)已经有了很大程度上的进步，但作为藏文信息处理的基础任务，其资源建设与技术发展仍处于初级阶段。见表3：

Table 3. Current status and application prospect of Tibetan semantic role tagging resources
表 3. 藏语语义角色标注资源建设现状与应用前景

方面	当前状况	主要挑战	发展前景
标注语料规模	约 8600 句(旦正才让, 2023)	覆盖领域有限，规模不足	扩大语料规模，增加领域多样性
标注体系	多种体系并存(2146 种角色)	缺乏统一标准	建立跨研究团队的标注标准
标注性能	F 值约 84%(规则+统计)	复杂句式处理不足	结合深度学习提升性能
应用场景	句法分析辅助、初步探索	实际应用案例有限	拓展到翻译、问答、信息抽取

5. 小结

藏语语义角色研究经过多年发展，已形成了较为完整的研究体系，在理论基础、体系构建、标注方法、资源建设等方面均取得了显著进展，并在多个领域展现出应用潜力。然而，面对当前存在的诸多挑战，未来需要在资源建设、方法创新、应用拓展等方面协同发力。通过加强多学科合作，推动标注体系标准化，不断探索新的技术方法和应用场景，藏语语义角色研究必将为藏语自然语言处理技术的发展注入新的活力，为传承和弘扬藏族文化、促进多语言信息交流与融合发挥重要作用。

参考文献

[1] 完么扎西, 嘎玛扎西, 头旦才让. 藏语动词语义角色的定义及标注标记集[J]. 高原科学研究, 2023, 7(4): 86-94.

[2] 珠杰, 仁青诺布, 春燕, 等. 论元角色的藏语语义角色标注研究[J]. 高原科学研究, 2018, 2(3): 85-96.

[3] 祁坤钰. 面向信息处理的藏语语义角色研究[J]. 西北民族大学学报(自然科学版), 2014, 35(4): 19-26.

[4] 旦正才让. 藏文语义角色计量研究[D]: [硕士学位论文]. 兰州: 西北民族大学, 2023.

[5] 龙从军, 康才峻, 李琳, 等. 基于多策略的藏语语义角色标注研究[J]. 中文信息学报, 2014, 28(5): 176-181.

[6] 旦正吉. 藏语语义块自动识别与标注方法研究[D]: [硕士学位论文]. 西宁: 青海师范大学, 2024.

[7] 何向真, 万福成, 于洪志, 吴玺宏. 基于藏语语义分析的机器翻译技术研究[J]. 计算机工程与应用, 2015, 51(15): 134-137, 173.

[8] 万福成, 马宁, 何向真. 融合事件特征及语义角色标注的藏文信息抽取技术[J]. 广西师范大学学报(自然科学版), 2018, 36(2): 18-23.