

基于 Yak-LLaMA-7B 领域微调的藏医 文本实体关系自动标注方法

仁欠扎西^{1,2} 多拉^{1,2} 安见才让³ 完玛道吉^{1,2} 冷本扎西^{*4}

(1.青海师范大学计算机学院 青海西宁 810016 2.青海师范大学藏语智能全国重点实验室 青海西宁 810008
3.青海民族大学智能科学与工程学院 青海西宁 810007 4.甘肃民族师范学院信息工程学院 甘肃甘南 747000)

摘要 针对藏医文本实体关系标注数据不足及人工标注工作量较大的问题,提出基于 Yak-LLaMA-7B 领域微调的实体关系自动标注方法。将藏医词典条目转换为指令数据,采用 LoRA 进行参数高效微调,并引入知识融合层与语义一致性正则化,以辅助生成实体关系三元组;对自动标注结果进行人工复核与筛选,形成结构化样本。模型性能评估实验表明, Yak-LLaMA-7B 的 F_1 值为 82.31%,较 TiLamb-7B、Tibetan-LLaMA2-7B 和 Tibetan-Alpaca-7B 分别提高 2.49、5.77 和 7.18 个百分点。量化对比结果显示,4 bit 量化使 F_1 值由 82.3% 降至 81.1%,下降 1.2 个百分点,显存占用由 18 GB 降至 8 GB,减少约 55.6%。结果表明,该方法可用于藏医文本实体关系的辅助标注,为藏医知识图谱构建提供结构化数据基础。

关键词 Yak-LLaMA-7B; LoRA; 实体关系抽取; 正则化机制; 知识融合

DOI:10.16249/j.cnki.2096-4617.2026.03.013

中图分类号 TP391.41

文献标识码 A

文章编号 2096-4617(2026)03-130-11

引言

藏医药是中华民族传统医学的重要组成部分,其知识体系蕴含大量未结构化的古籍文献与专业术语,如药材属性、方剂配伍和病症关联。对上述知识进行系统化整理是藏医药文献研究的关键任务^[1]。已有藏医药古籍数字化及标准体系建设研究^[2],为文献的数字化保存和利用奠定了基础,但将数字化文本进一步转化为可计算的实体关系数据,仍需明确实体边界、统一关系类型并控制标注质量。大语言模型 (Large Language Model, LLM) 通过预训练学习文本表示,并借助领域微调适配具体任务^[3],为上述结构化处理提供了技术路径。

针对藏医领域标注数据不足与人工标注工作量大的问题,本文利用藏语大模型辅助完成实体关系的自动标注。大语言模型通过大规模语料训练,可捕捉上下文信息与语言规律,目前已被广泛用于辅助写

收稿日期:2025-12-02

基金项目:青海省自然科学基金项目(2025-ZJ-922Q);藏语智能全国重点实验室课题(2025-SKL-002)。

第一作者简介:仁欠扎西,男,藏族,甘肃甘南人,青海师范大学计算机学院博士研究生,主要研究方向为藏医药信息处理。

通讯作者简介:冷本扎西,男,藏族,青海同仁人,甘肃民族师范学院信息工程学院副教授,主要研究方向为藏语计算语言学、藏医药信息处理。

作、文本润色和故事生成等任务^[4-6]。随着相关研究的推进,参数高效微调与低资源语言知识融合为藏语等少数民族语言的领域适配提供了新的思路。在实体关系抽取中,传统方法多依赖规则模板或有监督学习,受到人工标注成本和领域数据规模的限制,因此需要结合领域知识与任务数据开展模型适配。

在生成模型辅助数据构建方面,Møller等^[7]比较了人工标注数据与大语言模型合成增强数据在分类任务中的表现,指出合成数据的作用受任务及类别分布影响;Zhao等^[8]提出Genixer模型,通过指令模板设计、模型训练以及数据生成与筛选,构建多模态指令数据;Kim等^[9]提出DataDream模型,利用少量真实样本对扩散模型进行LoRA微调,再生成用于图像分类的训练数据。上述研究为利用生成模型补充任务数据提供了参考,同时也表明了生成数据需要结合具体任务进行质量评估。

在模型适配与医学应用方面,Liu等^[10]提出了P-Tuningv2,为较少可训练参数适配的下游任务提供了方法参考;Team等^[11]提出的ChatGLM系列为中文语言理解和生成研究提供了开源模型基础;Xiong等^[12]在ChatGLM-6B上开展医学对话微调,构建DoctorGLM;HuatuoGPT模型^[13]与BianQue模型^[14]进一步探索了医学问答、模拟问诊和多轮医疗对话;也有学者将LLaMA^[15]迁移至中文医疗知识领域,构建了HuaTuo(BenTsao)^[16]大模型;张君冬等^[17]成功构建了知识引导的中医古籍对话数据集,并训练了Huang-Di大模型。上述研究为医学领域的模型适配提供了参考,但相关语料与任务主要面向中文,仍需针对藏文的语言特点及藏医术语进行适配。

在藏文数据资源建设方面,尕藏扎西等^[18]开展了句法成分标注数据集研究;张洪溪等^[19,20]构建了结合模板的阅读理解数据集与动词结尾型问答数据集;官却才让等^[21]利用数字实体特征生成藏文问答数据。上述工作为藏文任务数据的自动构建提供了参考,也为藏语大模型的训练与评测积累了数据基础。

在藏语大模型研究方面,朱孟笑等^[22]通过翻译方法构建藏语指令数据集,以补充人工标注数据,最终构建了包含12个子任务的384 K条藏语指令数据,并对LLaMA-2和Baichuan-2进行了指令微调,该实验结果表明,微调后模型的藏语理解与生成能力得到提升,验证了藏语指令数据集的有效性;杨毛加等^[23]开展了基于参数高效微调的藏文大模型研究,通过LoRA等技术在有限藏文语料上实现模型适配,为藏医术语处理提供了基础模型支持;庄文浩等^[24]以LLaMA2-7B为基座模型,扩充词表,增加约30 000个藏文token,有效提高了藏文编码效率与语义理解能力,并通过增量预训练得到藏文大语言模型基座TiLamb。

现有研究为藏语大模型的语言能力与任务适配提供了基础,但面向藏医文本的实体边界识别、关系类型约束及标注结果质量控制仍需进一步探索。藏文复合术语、格助词结构及医学实体关系表达具有较强的专业性,直接应用通用模型可能出现术语切分不完整、实体边界不稳定和关系类型错配等问题。为此,本文提出基于Yak-LLaMA-7B领域微调的藏医文本实体关系自动标注方法,围绕基座模型选择、藏医词典数据预处理与指令转换、领域适配及三元组生成开展研究,并通过模型对比、消融实验与量化分析,评估方法的抽取性能和资源需求,为后续知识图谱构建、信息检索及藏医智能问答提供结构化数据基础。

1 研究方法

藏医领域数据稀缺与标注成本过高成为技术落地的核心瓶颈。例如,藏医药知识散见于古籍文献,缺乏公开的结构化标注实体关系数据。传统数据增强方法(如规则替换、回译等)难以捕捉领域知识的复杂性,而依赖人工标注的模式在成本与效率层面均难以为继。为此,本文通过LoRA参数高效微调技术,构建了藏医领域指令数据集与微调模型,以实现藏医文本中实体关系三元组的自动标注。

本研究首先从藏文支持、参数规模与编码方式等方面阐述基座模型的选择依据;其次,依据藏医词典确定实体类型和关系标签,并将条目字段转换为包含任务指令、输入文本和目标三元组的指令样本;随

后,以 Yak-LLaMA-7B 为主要基座,结合 LoRA 参数微调、藏医知识融合层与语义一致性正则化,实现从输入文本到结构化三元组的生成。生成结果经人工复核与筛选后形成标注数据,方法效果通过模型对比、消融实验和量化分析进行综合评价。

1.1 自动标注总体框架与基座模型选择

本文的自动标注流程由词典数据预处理、指令样本构建、模型领域微调 and 标注结果复核组成,总体框架如图 1 所示。具体流程包括 4 个步骤:首先,收集藏医词典条目并统一字段与文本格式;其次,将条目内容映射为包含任务指令、输入文本和目标三元组的训练样本;再次,以 Yak-LLaMA-7B 为主要基座模型,采用 LoRA 开展领域微调,并引入知识融合与语义一致性约束;最后,利用微调模型生成候选三元组,经人工复核和筛选后形成标注数据。

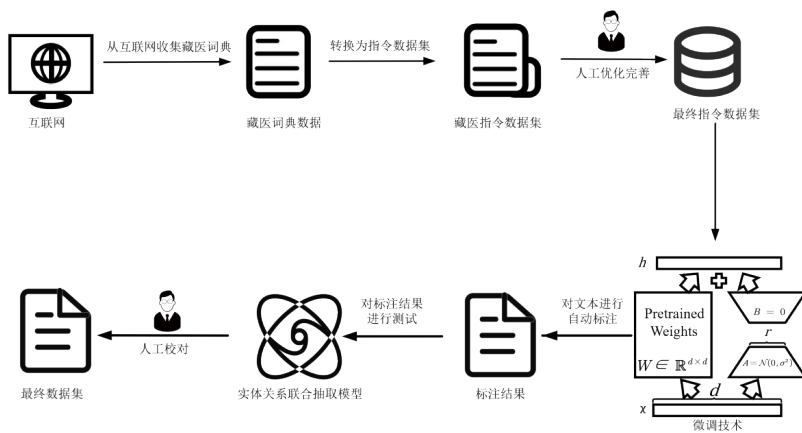


图 1 实体关系自动标注方法的总体框架

Fig.1 Overall workflow of the automatic entity-relationship annotation method

在模型比较实验中,分别以 TiLamb-7B、Yak-LLaMA-7B、Tibetan-LLaMA2-7B 和 Tibetan-Alpaca-7B 的公开权重为初始参数,使用藏医指令数据开展 LoRA 适配。4 个模型间的比较主要用于观察基座模型的差异,而 Yak-LLaMA-7B 则主要用以分析知识融合、语义约束消融及量化分析。

模型选择主要考虑藏文支持情况、参数规模和文本编码方式。4 个候选模型均为 7B 参数级模型,但基座权重、词表与训练过程并不完全相同。因此,相近的参数规模只能控制部分比较条件,不能据此认定模型具有相同的语言能力。本文结合表 1 的编码示例与 2.4 节的抽取实验分析其任务表现。

预训练与指令适配数据的差异主要涉及语料来源、领域覆盖和训练阶段。TiLamb 采用词表扩充与增量预训练,藏语指令数据构建及参数高效适配研究则分别关注任务样本与训练方式。由于本文未统计 4 个模型的预训练语料,因此未依据传统文献或网络文本的比例推断其领域能力,而是通过相同任务下的实验进行模型间比较。

分词结果用于观察文本的编码粒度与序列长度,不能单独衡量模型的语言理解能力。本文选取藏药描述文本“བྱུར་བྱུམ་ནི་རྩི་མང་ར་དང་ཁ་བ། རྩི་ཤང་བ་དང་། མཆིན་ནད་ཀྱན་དང་མཆིན་ཁྱག་ཁྱུས་པར་ཕན། ”(藏红花味甘、苦,性凉,主治肝病及肝血增盛)为例,比较 4 个模型的分词器输出能力。如表 1 所示,Yak-LLaMA-7B 生成 15 个 token,并将“བྱུར་བྱུམ་”“ཤང་བ་”“མཆིན་ནད་”分别保留在单个 token 中;Tibetan-LLaMA2-7B 和 Tibetan-Alpaca-7B 分别生成 18 个和 19 个 token,其中“བྱུར་བྱུམ་”被切分为两个独立的 token;TiLamb-7B 生成 30 个 token,包含较细的字符片段。上述结果仅反映该示例的编码差异。token 数量较少可缩短该文本的输入序列,但不足以证明术语语义保存程度或下游抽取能力更高。综上,该文本在 Yak-LLaMA-7B 下具有较短的 token 序列,且部分术语未被进一步拆分。因此,本文以其作为主要基座模型,并保留其余 3 个模型开展对比实验。

表1 分词结果比较
Table 1 Comparison of segmentation results

模型	分词结果	长度
TiLamb-7B	“__ག”“རུ”“གུམ”“ནི”“རྩམ”“ངར”“དང་ཁ”“བ”“།” “__ ང”“ོ”“ཐོ”“གང”“བ་དང་”“།” “__ མ”“ཆ”“ི”“ན”“ནད” “ཀུན”“དང”“མ”“ཆ”“ི”“ན”“ཁྲག”“རྩམ”“པར”“པན།”	30
Tibetan-Alpaca-7B	“__གུར”“གུམ་ནི”“རྩམ”“མངར”“དང”“ཁ”“བ།” “__ རོ་ཐོ”“གང་བ”“དང།” “__ མཆེན”“ནད”“ཀུན”“དང”“མཆེན” “ཁྲག”“རྩམ”“པར་པན།”	19
Tibetan-LLaMA2-7B	“__གུར”“གུམ་ནི”“རྩམ”“མངར”“ དང ”“ཁ”“བ།” “__ རོ་ཐོ”“གང་བ”“དང།” “__ མཆེན”“ནད”“ཀུན”“དང”“མཆེན” “ཁྲག”“རྩམ”“པར་པན།”	18
Yak-LLaMA-7B	“__གུར་གུམ”“ནི”“རྩམ”“མངར”“དང”“ཁ་བ།” “__ རོ་ཐོ”“གང་བ”“དང།” “__ མཆེན་ནད”“ཀུན”“དང”“མཆེན་ཁྲག”“རྩམ” པར”“པན”	15

1.2 藏医词典收集与预处理

1.2.1 实体

实体关系抽取作为藏医药信息抽取的核心子任务,其主要目标是从非结构化或半结构化的藏医文本中识别并结构化呈现具有医学语义关联的实体及其关系网络。实体作为知识图谱构建的基本要素,是指文本中承载特定医药学语义信息并参与医疗逻辑表达的语言单位。在通用领域研究中,实体通常包含人物类、组织类、地理类、时间类及数值类等基础类型。

结合藏医词典中的方剂、药物、病症、功效、文献来源和用法信息,本文设置6类实体,用于统一指令样本及标注结果中的实体类型。实体类型的确定以条目内容和具体上下文为依据,药性等描述性字段通过相应关系与药物关联。各类实体的具体含义如下:

- (1)方剂实体:指由多种药物配伍组成的制剂名称,如“ཙང་ཞི་ཉེར་གཅིག”(石膏二十一味)。
- (2)药物实体:指词典中记录的单味药或药材名称,包括植物、动物和矿物来源的药材,如“ཤིང་མངར”(甘草)、“བླ་ཤང”(竹黄)等。
- (3)病症实体:指文本记载的疾病、症状或病理状态名称,如“རྩུང་ནད”(隆病)。标注时依据原文确定其边界及与其他实体的关联。
- (4)功效实体:指原文描述的药物或方剂作用。该类实体依据词典中其功效字段及上下文进行标注,用于表达药物或方剂与相应作用之间的关系。
- (5)来源实体:指方剂、药物或相关知识所依据的文献,如“བླ་ཤང་ཞི”(《四部医典》)、“མན་ངག་རིན་ཆེན་འབྱུང་གནས”(《秘诀宝源》)。
- (6)用法实体:指原文记载的给药时间、频率、数量或方式,如“ཉིན་རེར་ཐེངས་གཅིག”(每日一次)。该类内容作为文本信息进行标注。

上述分类为实体边界识别和关系标注提供统一的类型约束,并为后续人工复核与结构化存储提供依据。

1.2.2 实体关系

实体关系用于表示文本中两个实体之间的语义关联。根据藏医词典中的组成、功效、药性、用法和来源等字段,本文设置11类关系标签,具体含义见表2。关系标注需同时考虑字段语义和文本上下文,以减少同一关联采用不同标签的情况。

表2中的关系标签覆盖组成、作用与文献来源等信息:“配方”“构成”和“药物”描述药材与方剂或制剂之间的联系;“功效”“益处”和“治疗”描述文本记载的作用或病症关联;“包含”和“来源”描述文献收录与溯源关系。其余标签用于表示引药、药性和用法信息。本研究标注结果严格采用表2所列标签,避免混用

表2 藏医实体关系表
Table 2 Entity relationship table of Tibetan medicine

实体关系	语义特征	实体关系	语义特征
配方	方剂与组成其单味药间的隶属关系	来源	方剂在原始文献中的溯源关系
引药	辅药与主方之间的关联	药性	药物的属性或特征
药物	药材与方剂的匹配关系	用法	施治中给药的数量关系
功效	方剂对特定病症的效应	治疗	疗法与特定疾病的对应关系
益处	药材对症状的缓解效应	构成	药材与衍生制剂之间的包含关系
包含	医学典籍对方剂的收录关系		

未定义的关系名称。

以“བཤེ་ཅུ་ཉེར་ལྷ།”(犀角二十五味)为例,条目中的组成字段可关联“ཅན་དན་དཀར་པོ།”(白檀香)、“ཨ་ཅུ་ར།”(诃子)等药物实体,功效或主治字段可关联相应作用与病症。抽取过程将这些条目信息转换为三元组,并依据表2的统一关系名称。此类结构化表达用于记录原文知识及其关联。

如图2所示,传统藏医词典数据以条目式结构呈现,其中包含“གྲུ་གུམ།”(藏红花)、“རྩ་བོ།”(性)、“རྩ་མངར་ཞིང་ལ་བ།”(味甘、苦)、“རྩ་བོ་བསེལ་བ།”(性凉)、“ཕན་ལུས།”(功效)、“སྤྲུལ་བཅོས།”(治疗)等关键字段。如,某一条目通过“བྱུག་གཟེར་སེལ་བ།”(散瘀止痛)、“སེམས་དྲངས་ཡི་ཟུག་སེལ་བ།”(清心解郁)等子项,可清晰呈现藏药功效与对应病症的关联。

为适配大模型指令微调需求,需将传统词典格式的藏医术语数据转换为标准化的“输入-输出”配对指令格式,如图3所示。原始数据采用传统藏医词典结构,以“གྲུ་གུམ།”(藏红花)为例,词典条目包含术语名称、性质描述(རྩ་བོ)、功能作用(ཕན་ལུས)和治疗适应症(སྤྲུལ་བཅོས)等字段信息。

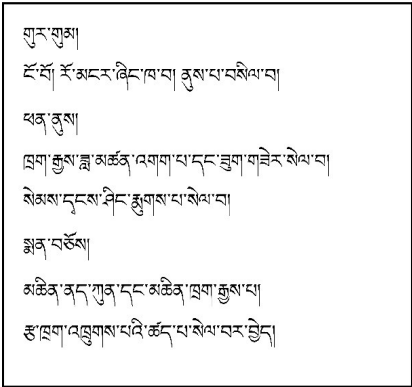


图2 藏医词典格式
Fig.2 The format of Tibetan medicine dictionary

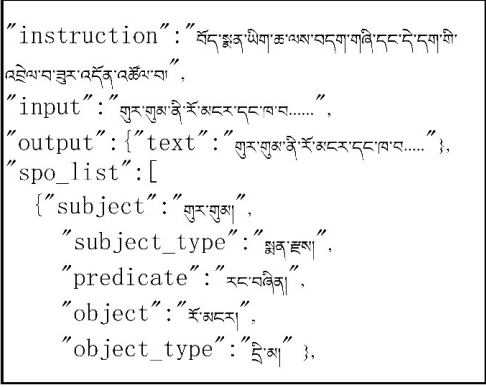


图3 指令数据集格式
Fig.3 The fromat of instruction dataset

通过自定义脚本将词典字段映射至指令模板,转换为结构化的指令数据集。转换后的标准格式包含3个核心组件:instruction 字段提供明确的任务指令,如“བོད་སྐད་ཡིག་ཆ་ལས་བདག་གཞི་དང་དེ་དག་གི་འབྲེལ་བ་བྱུར་འདོན་ཅིང་བ།”(从藏医文献中抽取实体及其关联关系);input 字段包含术语的核心信息,整合了术语名称、性质特征、功能作用及治疗适应症等内容;output 字段则为结构化的实体关系标注结果,采用三元组形式(头实体,关系类型,尾实体)表示实体间的语义关联。转换过程中,实体类型涵盖“སྤྲུལ་བཅོས།”(药物)、“ནད་རྟགས།”(病症)等多个类别,关系类型包括“རང་བཞིན།”(性质)、“ཕན་ལུས།”(功效)等关系。以“གྲུ་གུམ།”为例,其药性信息“རྩ་མངར་ཞིང་ལ་བ།”(味甘且苦)被标注为药物实体与味觉属性的关联关系。

词典字段完成映射后,每条样本均以 instruction 说明抽取任务,以 input 提供条目描述,以 output 中的

spo_list 保存实体关系三元组。字段映射则负责将已有词典知识转化为监督样本,模型则学习从描述文本生成相应三元组。形式化表达公式如下:

$$T = f_{map}(E, A, R) \quad (1)$$

式中, E, A, R 分别为实体集合、属性与关系类别, f_{map} 表示词典字段到模板的自动映射函数。

1.3 藏医领域适配与实体关系生成

1.3.1 LoRA 参数微调方法

全参数微调需要更新预训练模型的全部参数,并存储相应梯度及优化器状态,资源开销随模型规模增大而增加。低秩适配(Low-Rank Adaptation, LoRA)通过冻结基座权重、训练低秩增量矩阵,以减少任务适配所需更新参数。

本文采用 LoRA 对 4 个候选藏文模型进行适配,并在注意力层的查询投影和数值投影模块中设置可训练的低秩矩阵。基座模型权重在训练中保持固定,参数更新由适配器承担。资源需求取决于适配位置、秩、序列长度及计算精度。

对于 Transformer 结构中的线性层,其权重矩阵为 $W_0 \in R^{d \times k}$,其中, d 为输出维度, k 为输入维度。LoRA 还添加了两个低秩分解的可训练矩阵 $B \in R^{d \times r}$ 和 $A \in R^{r \times k}$,其中, r 为预定秩。在前向传播过程中, W_0 与 ΔW 均需乘以相同的输入 x 并相加,公式如下:

$$h = W_0 x + \Delta W_x = W_0 x + B A x, B \in R^{d \times r}, A \in R^{r \times k} \quad (2)$$

在 LoRA 训练过程中, W_0 保持固定,仅更新低秩矩阵 A 和 B 。该机制减少了需要训练和存储优化器状态的参数数量,为计算资源受限条件下的领域适配提供可行的方法。本文利用已有藏医指令数据进行适配,具体数据规模和硬件条件见 2.1 节和 2.2 节。

本文采用藏医指令样本训练 LoRA 适配器,使模型学习由条目描述到关系三元组的映射。冻结基座权重可以减少更新范围,但不能据此认定模型避免了过拟合或完整保留了原有能力。本文通过后续抽取实验比较领域适配结果。

1.3.2 藏医知识融合层

LoRA 负责以较少参数完成任务适配,其中的知识融合层则用于向模型提供藏医实体信息。本文在 LoRA 框架中引入知识融合层(Knowledge Fusion Layer),以藏医词典为知识来源,将检索到的实体嵌入 K_{domain} ,并与当前文本的隐藏表示进行注意力融合。融合后的表示参与后续实体关系生成,其计算公式如下:

$$H' = H + \alpha \cdot \text{Attn}(H, K_{domain}) \quad (3)$$

式中, H 为文本隐藏表示, K_{domain} 为藏医领域实体知识嵌入, α 为融合系数。注意力项依据文本表示对实体知识进行聚合,并通过残差形式与原表示相加。实体检索采用余弦相似度匹配,从藏医知识库中选择与输入相关的实体信息,再将其嵌入至融合层。该过程为生成三元组补充外部知识信息。

训练阶段,输入的任务指令与文本首先经模型编码,检索得到的实体嵌入通过知识融合层与文本隐藏表示进行融合,LoRA 适配器则用于学习任务所需的参数增量。模型以结构化三元组为目标生成序列,并依据训练损失更新可训练参数。推理阶段,模型输出候选三元组,供后续人工复核。知识融合层的作用通过 2.5 节的消融结果进行分析。

1.3.3 语义一致性正则化机制

模型生成的三元组可能在关系类型或实体内容上偏离输入文本。为进一步约束输入文本与输出三元组在表示空间中的相关性,本文引入语义一致性正则化(Semantic Consistency Regularization)。在交叉熵损失 L_{CE} 之外,以输入和输出语义向量的余弦相似度构建约束项。训练目标公式如下:

$$L=L_{CE}+\lambda(1-\cos(E_{input},E_{output}))$$

(4)

式中, E_{input} 为输入文本的语义向量, E_{output} 为输出三元组的语义嵌入, λ 为平衡系数。该项在训练中通过最小化输入输出语义距离,使模型保持语义连贯与逻辑一致。

使用 Yak-LLaMA 的中间层表示作为句向量表示,通过均值池化获得 E_{input} 与 E_{output} 。语义相似度由余弦距离计算得出,并在反向传播阶段联合更新 LoRA 参数。该方法有效防止了模型因指令式生成而产生语义断裂现象。

通过在验证集上的可视化实验发现,引入语义一致性约束后,模型在自动标注藏医文本时生成的三元组语义更加贴合原文含义。如,句子“གྲུ་གུམ་ནི་སེམས་དྲངས་ཡི་ཐུག་སེལ་བའི་ཕན་ཐྱས་ཡོད།”(藏红花具有清心解郁之功效),在未加入语义一致性约束的模型在自动标注时常输出:“གྲུ་གུམ་”“ནད་སྤྱུམ་”“གཏོང་བ།”“མགོ་ན།”(藏红花-治疗-头痛),虽然结构形式正确,但与原句中表达的“具有功效”语义层级不符。加入语义一致性约束后,模型输出:“གྲུ་གུམ་”“ཕན་ཐྱས་ཡོད་པ།”“སེམས་དྲངས་ཡི་ཐུག་སེལ།”(藏红花-具有功效-清心解郁),与原文语义完全一致。定量分析表明,引入语义一致性约束后,自动标注结果的关系类型准确率与语义一致性得分平均提升约 9%,显著减少了关系类型错配与语义偏移现象。

2 实验设置与结果分析

2.1 实验数据

指令微调数据由藏医词典条目转换得到,共计 2 872 条,采用 instruction、input 和 output 字段组织。模型对藏医文本生成候选实体关系标注后,经人工复核与筛选得到 24 462 条样本,涵盖药物、功效和病症等信息。该样本集按 8:1:1 划分为训练集、验证集和测试集,用于后续实验。

2.2 训练配置

模型比较包括 TiLamb-7B、Tibetan-LLaMA2-7B、Yak-LLaMA-7B 和 Tibetan-Alpaca-7B,均以各自公开权重为初始参数。训练采用 LoRA,将适配器设置于注意力机制的 q_proj 和 v_proj 模块。训练设置采用 4 bit 加载以节省显存,并使用 fp16 计算;分别记录权重存储精度与训练计算精度。主要超参数见表 3,量化对比结果见表 7。

表 3 模型参数设置
Table 3 Parameter settings of the models

设置	Yak-LLaMA-7B	TiLamb-7B	Tibetan-LLaMA2-7B	Tibetan-Alpaca-7B
训练轮次	3	3	3	3
有效批量大小	16	16	16	16
学习率	2e-4	2e-4	2e-4	2e-4
LoRA	“q_proj”	“q_proj”	“q_proj”	“q_proj”
权重	“v_proj”	“v_proj”	“v_proj”	“v_proj”
计算精度	fp16	fp16	fp16	fp16
LoRA 秩	8	8	8	8
训练时长	3.3 h	2.8 h	3.2 h	3.1 h

预处理将指令样本组织为“问:{instruction}{input}\n 答:{output}”序列,其中\n 表示换行。使用各模型对应的 Llama Tokenizer 进行编码,序列长度上限设为 256 个 token,超过上限的序列截断,不足上限的序列按训练设置填充。训练采用因果语言模型目标,以 input_ids 构造 labels,并使用非掩码语言模式的 DataCollator 处理批数据。单卡训练使用 24 GB 显存的 NVIDIA GPU;每设备批次为 2,梯度累积步数为 8,有效批次为 16,共训练 3 个 epoch,学习率为 2e-4,计算精度为 fp16。每 50 步记录日志,每 500 步保存检查点,

最多保留 2 个检查点,主要超参数见表 3。

2.3 模型训练损失对比

本文记录 4 个模型在相同训练语料与主要超参数设置下的训练损失,用于观察优化过程。各模型训练 3 个 epoch,损失曲线如图 4 所示。不同分词器对应的序列与预测单元可能不同,因此跨模型损失值仅作为优化过程的参考,不能直接替代实体关系抽取指标。

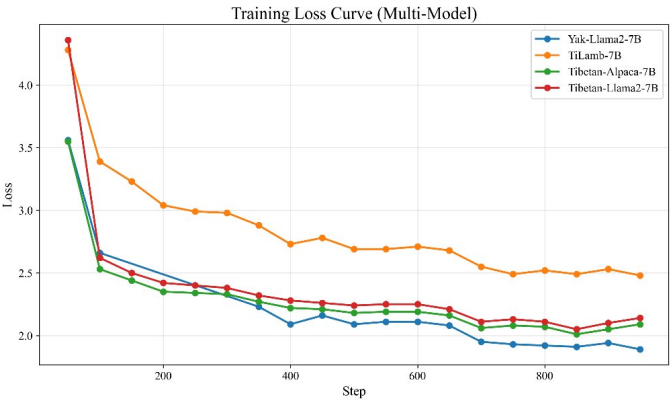


图 4 训练损失对比
Fig.4 Comparison of training loss

图 4 中,4 个模型的训练损失值均随着训练步数的增加呈下降趋势,并且未出现明显过拟合现象,表明 LoRA 微调在该任务中能够有效学习领域知识。具体而言:Yak-Llama-7B 在约 600 步时损失值迅速下降并进入稳定区间,最终收敛至 2.1 左右,说明其对“药材病症-功效”类关系的建模能力较强;TiLamb-7B 与 Tibetan-Llama2-7B 的损失下降曲线较为平缓,在 800 步左右收敛至 2.3~2.4 区间;Tibetan-Alpaca-7B 的整体下降速度最慢,最终损失维持在 2.5 左右,表明其在专业领域知识适配方面存在一定劣势。

在实验记录中,各模型的训练目标均有所下降。是否具有较好的泛化表现,应结合独立评测数据上的精确率、召回率和 F1 值分析,相关结果见 2.4 节。

2.4 实体关系联合抽取性能评估

实验比较了 4 个模型在藏医文本实体关系联合抽取任务中的结果。评价对象为生成的实体关系三元组,采用精确率、召回率与 F_1 值描述抽取结果的正确性和覆盖情况。

本文采用精确率(Precision)、召回率(Recall)与 F_1 值(F_1 -score)作为核心指标,具体计算公式如下:

$$Precision = \frac{TP}{TP + FP}$$
 (5)

$$Recall = \frac{TP}{TP + FN}$$
 (6)

$$F_1 = 2 \times Precision \times \frac{Recall}{Precision + Recall}$$
 (7)

式中,TP(True Positive)为正确抽取的实体关系三元组数量,FP(False Positive)为错误抽取的三元组数量,FN(False Negative)为未能识别出的三元组数量。Precision 反映了抽取结果的准确性,Recall 体现了覆盖完整性, F_1 值则是二者的调和平均,用于综合衡量模型性能。

表 4 列出了 4 个模型的实体关系抽取结果。其中,Yak-LLaMA-7B 的精确率、召回率和 F_1 值分别为 83.12%、81.53% 和 82.31%;TiLamb-7B、Tibetan-LLaMA2-7B 和 Tibetan-Alpaca-7B 的 F_1 值分别为 79.82%、76.54% 和 75.13%。Yak-LLaMA-7B 相对另外 3 个模型的 F_1 值分别提高了 2.49、5.77 和 7.18 个百分点。上述差异反映了 4 个微调模型在本文数据及实验配置下的整体抽取表现。在所比较的 4 个模型中,Yak-LLaMA-7B 取得了较高的抽取指标,但是该结果仅支持在本文任务范围内继续分析其模块作用及量化效

果,尚不足以证明方法在其他语料、语言或应用场景中的普适性。且针对长文本、复杂方剂及罕见关系类型等差异,仍需按样本特征和关系类别开展进一步评测。

表4 模型实体关系抽取结果比较
Table 4 Comparison of entity-relationship extraction results of the models

模型	$P(\%)$	$R(\%)$	$F_1(\%)$
Yak-LLaMA-7B	83.12	81.53	82.31
TiLamb-7B	80.24	79.51	79.82
Tibetan-LLaMA2-7B	77.13	76.02	76.54
Tibetan-Alpaca-7B	75.81	74.52	75.13

为验证模型在实际应用中的表现,本文从藏医典籍中选取典型语句进行实体关系抽取实例分析。以输入文本“ $\text{གྲུར་གྲུམ་ནི་སེམས་དྲུངས་ཡི་ཐུག་སེལ་བའི་པན་རྒྱས་ཡོད།}$ ”(藏红花具有清心解郁之功效)为例,其在各模型中的输出结果如表5所示。

表5 模型标注实例分析
Table 5 Annotation case analysis of the models

模型	输出三元组	评价
Yak-LLaMA-7B	$(\text{གྲུར་གྲུམ}, \text{པན་རྒྱས}, \text{སེམས་དྲུངས་ཡི་ཐུག་སེལ})$	语义准确,关系类型正确
TiLamb-7B	$(\text{གྲུར་གྲུམ}, \text{ནད་ཐྱུག་གཏོང་བ}, \text{མགོན})$	关系类型错误,语义偏离
Tibetan-LLaMA2-7B	$(\text{གྲུར་གྲུམ}, \text{གྲུབ་པ}, \text{སེམས་དྲུངས་ཡི་ཐུག་སེལ})$	语义正确,但关系错误
Tibetan-Alpaca-7B	$(\text{གྲུར་གྲུམ}, \text{བཅོས་པ}, \text{སེམས་དྲུངས})$	关系类型不精确,语义不完整

2.5 消融实验

为分析知识融合层与语义一致性正则化对Yak-LLaMA-7B抽取结果的影响,设置完整模型、去除知识融合层和去除语义一致性正则化等3种实验配置。在相同的数据划分与主要超参数设置下比较 F_1 值,结果见表6。

表6 消融实验
Table 6 Ablation experiment

实验设置	$F_1(\%)$	变化(百分点)
完整模型	82.3	—
去除知识融合层	80.4	↓ 1.9
去除语义一致性正则	81.0	↓ 1.3

由表6可知,去除知识融合层后,模型的 F_1 值降低1.9个百分点,说明该模块在一定程度上提升了模型对领域知识的理解能力。知识融合层通过引入外部藏医领域实体知识嵌入,使模型在注意力计算中能够同时考虑文本上下文与实体语义信息,从而增强模型对“药材-功效-病症”等复杂关系的识别能力。同时,移除语义一致性正则化后,模型性能下降1.3个百分点。表明该机制有效减少了生成三元组过程中的语义偏离,使输出结果与原始文本语义保持较好的对应关系。

2.6 量化与轻量化效果

为进一步分析降低权重存储精度对抽取指标和资源占用的影响,本文比较Yak-LLaMA-7B的FP16模型与4 bit量化模型。记录精确率、召回率、 F_1 值、显存占用及相对推理速度,量化对比结果如表7所示。

表7 Yak-LLaMA-7B模型量化对比
Table 7 Quantization comparison of Yak-LLaMA-7B Model

模型	$P(\%)$	$R(\%)$	$F_1(\%)$	显存占用(GB)	推理速度提升
原始FP16模型	83.1	81.5	82.3	18	—
4 bit量化模型	82.0	80.2	81.1	8	+40%

如表7所示,4 bit量化后显存占用由18 GB降至8 GB,减少10 GB,降幅约为55.6%;表中记录的相对推理速度提升为40%。精确率由83.1%降至82.0%,召回率由81.5%降至80.2%, F_1 值由82.3%降至81.1%,上述分别下降1.1、1.3和1.2个百分点。

上述结果反映了本次量化对比中显存占用下降与抽取指标变化的关系,为选择模型部署配置提供参考。推理速度受硬件、输入长度、批次及测量方式影响,不能将该提升比例直接推广至其他设备。

结语

本研究基于Yak-LLaMA-7B实现了藏医文本的自动实体关系标注,融合LoRA参数高效微调、知识融合层与语义一致性约束,实现藏医领域实体关系的自动抽取与标注。通过与TiLamb-7B、Tibetan-LLaMA2-7B和Tibetan-Alpaca-7B等7B参数级藏文大模型进行对比实验发现,Yak-LLaMA-7B在分词完整性、训练收敛速度和实体关系抽取方面均表现出更好的性能,说明其更适用于藏医专业文本的实体关系自动标注任务。进一步采用4 bit量化后该模型仍然能保持较高的性能,验证了该模型在低算力场景中的实用性。未来工作将结合RAG检索增强与专家反馈,构建更完整的藏医知识服务体系。

参考文献

- [1] 罗婕,王汝珊,宋鑫晨,等.藏医药文献研究现状及传承思考[J].中华中医药杂志,2022,37(1):60-64.
- [2] 刘晓建,季拥政.藏医药古籍文献数字化标准体系示范建设[J].数字图书馆论坛,2021(2):27-33.
- [3] 吴春志,赵玉龙,刘鑫,等.大语言模型微调方法研究综述[J].中文信息学报,2025,39(2):1-26.
- [4] Pereira C A, Komarlu T, Mobeirek W. The future of AI-assisted writing[PP/OL]. V1. arXiv (2023-06-29) [2026-09-08]. <https://arxiv.org/abs/2306.16641>. DOI:10.48550/arXiv.2306.16641.
- [5] Shi S, Zhao E, Bi W, et al. Effidit: an assistant for improving writing efficiency[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: Volume 3, System Demonstrations. Toronto: Association for Computational Linguistics, 2023:508-515.
- [6] Yuan A, Coenen A, Reif E, et al. Wordcraft: story writing with large language models[C]//Proceedings of the 27th International Conference on Intelligent User Interfaces. New York: Association for Computing Machinery, 2022:841-852.
- [7] Møller A G, Pera A, Dalsgaard J, et al. The parrot dilemma: human-labeled vs. LLM-augmented data in classification tasks[C]//Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. St. Julian's, Malta: Association for Computational Linguistics, 2024:179-192.
- [8] Zhao H H, Zhou P, Shou M Z. GENIXER: empowering multimodal large language models as a powerful data generator[C]//Computer Vision-ECCV 2024. Cham: Springer, 2025:129-147.
- [9] Kim J M, Bader J, Alaniz S, et al. DataDream: few-shot guided dataset generation[C]//Computer Vision-ECCV 2024. Cham: Springer, 2025:252-268.
- [10] Liu X, Ji K, Fu Y, et al. P-Tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks[PP/OL]. V3. arXiv (2022-03-20) [2026-09-08]. <https://arxiv.org/abs/2110.07602>. DOI:10.48550/arXiv.2110.07602.
- [11] Team G L M, Zeng A, Xu B, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 All Tools[PP/OL]. V2. arXiv (2024-07-30) [2026-09-08]. <https://arxiv.org/abs/2406.12793>. DOI:10.48550/arXiv.2406.12793.
- [12] Xiong H L, Wang S, Zhu Y T, et al. DoctorGLM: fine-tuning your Chinese doctor is not a herculean task[PP/OL]. V2. arXiv (2023-04-17) [2026-09-08]. <https://arxiv.org/abs/2304.01097>. DOI:10.48550/arXiv.2304.01097.
- [13] Zhang H B, Chen J Y, Jiang F, et al. HuatuoGPT, towards taming language model to be a doctor[C]//Findings of the Association

for Computational Linguistics: EMNLP 2023. Singapore: Association for Computational Linguistics, 2023:10859–10885.

[14] Chen Y R, Wang Z Y, Xing X F, et al. BianQue: balancing the questioning and suggestion ability of health LLMs with multi-turn health conversations polished by ChatGPT[PP/OL]. V2. arXiv (2023–12–04) [2026–09–08]. <https://arxiv.org/abs/2310.15896>. DOI:10.48550/arXiv.2310.15896.

[15] Touvron H, Lavril T, Izacard G, et al. LLaMA: open and efficient foundation language models[PP/OL]. V1. arXiv (2023–02–27) [2026–09–08]. <https://arxiv.org/abs/2302.13971>. DOI:10.48550/arXiv.2302.13971.

[16] Wang H C, Liu C, Xi N W, et al. HuaTuo: tuning LLaMA model with Chinese medical knowledge[PP/OL]. V1. arXiv (2023–04–14) [2026–09–08]. <https://arxiv.org/abs/2304.06975>. DOI:10.48550/arXiv.2304.06975.

[17] 张君冬,杨松桦,刘江峰,等. AIGC赋能中医古籍活化: Huang-Di 大模型的构建[J]. 图书馆论坛, 2024, 44(10): 103–112.

[18] 尔藏扎西,冷本扎西,多拉. 基于格序列的藏语句法成分标注数据集构建方法研究[J]. 高原科学研究, 2023, 7(4): 95–105.

[19] 张洪溪,才智杰. 结合模板的藏文机器阅读理解数据集自动构建方法[J]. 计算机工程与应用, 2026, 62(3): 198–206.

[20] 张洪溪,才智杰. 面向自动问答的藏文动词结尾型数据集构建[J]. 智能系统学报, 2025, 20(5): 1207–1216.

[21] 官却才让,杨毛加,柔特,等. 基于数字实体特征的藏文问答数据集构建[J]. 中文信息学报, 2025, 39(3): 59–65.

[22] 朱孟笑,沙九,冯冲. 面向大语言模型的藏语指令数据集构建[J]. 中文信息学报, 2024, 38(12): 83–96.

[23] 杨毛加,柔特,才智杰,等. 基于参数高效微调的藏文大模型研究[J]. 中文信息学报, 2024, 38(12): 106–115.

[24] 庄文浩,孙媛,赵小兵. TiLamb: 基于增量预训练的藏文大语言模型[C]//第23届中国计算语言学大会论文集: 卷1, 主会论文. 太原: 中国中文信息学会, 2024: 254–267.

Study on an Automatic Entity–Relationship Annotation Method for Tibetan Medicine Texts Based on Domain Fine–Tuning of Yak–LLaMA–7B

Renqian–Zhaxi^{1,2} Duola^{1,2} Anjian–Cairang³ Wanma–Daoji^{1,2} Lengben–Zhaxi^{*4}

(1. College of Computer Science and Technology, Qinghai Normal University, Xining 810016, China;

2. State Key Laboratory of Tibetan Intelligen, Qinghai Normal University, Xining 810008, China;

3. School of Intelligent Science and Engineering, Qinghai Minzu University, Xining 810007, China;

4. School of Information Engineering, Gansu Minzu Normal University, Gannan 747000, China)

Abstract: To address the issues of insufficient annotated entity–relationship data and the substantial labor demands of manual annotation in Tibetan medicine texts, this study proposes an automatic annotation method based on domain–specific fine–tuning of Yak–LLaMA–7B. Entries from the Tibetan medicine dictionary were converted into instruction datasets, and Low–Rank Adaptation(LoRA) was used for efficient parameter fine–tuning. Furthermore, we incorporated a knowledge fusion layer and semantic consistency regularization to generate entity–relationship triples. The automatically annotated results were manually reviewed and filtered to create a structured dataset. Model performance evaluation experiments showed that Yak–LLaMA–7B attained an F_1 score of 82.31%, representing improvements of 2.49, 5.77, and 7.18 percentage points over TiLamb–7B, Tibetan–LLaMA2–7B, and Tibetan–Alpaca–7B, respectively. Quantization comparison results show that 4–bit quantization led to a modest decrease in F1 score from 82.3% to 81.1%(a reduction of 1.2 percentage points), while concurrently reducing GPU memory consumption from 18 GB to 8 GB—an approximately 55.6% reduction. These results indicate that the method can be used for auxiliary annotation of entity–relationships in Tibetan medicine texts, providing a structured data foundation for the construction of a Tibetan medicine knowledge graph.

Keywords: Yak–LLaMA–7B, LoRA, entity–relationship extraction, semantic consistency regularization mechanism, knowledge fusion

[责任编辑:索郎桑姆]