智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向自动问答的藏文动词结尾型数据集构建

张洪溪, 才智杰

引用本文:
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.
ZHANG Hongxi, CAI Zhijie. Construction of a Tibetan verb-ending type dataset for automatic question answering[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(5): 1207-1216.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202410002

DOI: 10.11992/tis.202410002
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20250623.1446.004

面向自动问答的藏文动词结尾型数据集构建

张洪溪[1,2], 才智杰[1,2]
(1. 青海师范大学 计算机学院, 青海 西宁 810016; 2. 藏语智能全国重点实验室, 青海 西宁 810008)

摘 要: 自动问答数据集是研究藏文自动问答技术的重要数据基础。文章针对藏文自动问答数据集匮乏的瓶颈问题, 在剖析英文、汉文和藏文自动问答数据集构建现状的基础上, 分析了藏文中出现频率最高的动词结尾型句子的问答结构特征, 通过构建句子和问句的模板, 设计了一种面向自动问答的藏文“动词结尾+位格助词”型数据集构建方案, 按照方案构建了面向自动问答的藏文数据集 TiQuAD_36414, 并采用平均意见得分 (mean opinion score, MOS) 方法, BiDAF(bidirectional attention flow)、RNet(gated self-matching networks) 和 QANet(question answering net) 模型的 F1 值和 EM(exact match) 值验证了数据集的有效性。实验数据表明, 本文构建的数据集 TiQuAD_36414 的质量良好。

关键词: 自然语言处理; 藏文; 自动问答; TiQuAD_36414 数据集; 问答模板; 动词; 位格助词; 有效性
中图分类号: TP391 **文献标志码**: A **文章编号**: 1673−4785(2025)05−1207−10

中文引用格式: 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.
英文引用格式: ZHANG Hongxi, CAI Zhijie. Construction of a Tibetan verb-ending type dataset for automatic question answering[J]. CAAI transactions on intelligent systems, 2025, 20(5): 1207–1216.

Construction of a Tibetan verb-ending type dataset for automatic question answering

ZHANG Hongxi[1,2], CAI Zhijie[1,2]
(1. College of Computer Science and Technology, Qinghai Normal University, Xining 810016, China; 2. The State Key Laboratory of Tibetan Intelligence, Xining 810008, China)

Abstract: The Tibetan automatic question answering (Q&A) dataset serves as a crucial data foundation for advancing research in Tibetan automatic Q&A technologies. To solve the problem of the lack of automatic Q&A datasets in Tibetan, this paper first examines the features of the most common verb-ending type sentences in Tibetan based on an analysis of the current status of automatic Q&A dataset construction in English, Chinese, and Tibetan. Then, this study constructs templates for sentences and questions and proposes a template-based method for building a Tibetan automatic Q&A dataset with “verb-ending + La case auxiliary word” sentences. Then, a new Tibetan automatic Q&A dataset (TiQuAD_36414) is generated according to this approach. Finally, the validity of this dataset is verified using the MOS(mean opinion score) method, along with the F1 and EM(exact match) scores of the BiDAF(bidirectional attention flow), RNet(Gated Self-Matching Networks), and QANet(question answering net) models. The experimental results show that the performance of the TiQuAD_36414 dataset is better than that of the baseline Tibetan Q&A dataset.

Keywords: natural language processing; Tibetan; automatic Q&A; TiQuAD_36414 dataset; Q&A template; verb; la case auxiliary word; effectiveness

近年来, 随着信息技术的快速发展, 网络信息量呈现爆炸式增长。面对海量数据信息, 准确、快速地获取所需知识成为了越来越严峻的挑战[1]。自动问答技术的引入为解决海量信息筛选、提取

收稿日期: 2024−10−02. **网络出版日期**: 2025−06−24.
基金项目: 国家自然科学基金项目 (61966031,61866032); 藏文信息处理教育部重点实验室项目 (2013-Z-Y17, 2014-Z-Y32, 2015-Z-Y03).
通信作者: 才智杰. E-mail: Czjqhsd@163.com.

和理解提供了高效的途径，因此自动问答作为人工智能领域的一个重要分支受到了越来越多的关注[2]。

随着深度学习技术的突破和大语言模型的发展，ChatGPT 等大规模预训练语言模型把自动问答系统推向了新的高度[3]。大语言模型在训练过程中需要大量的数据以及足够的计算资源进行模型参数的优化和学习，然而藏文自动问答缺乏充足的训练数据和相关的语言理解技术，导致大语言模型性能表现较差。与汉文和英文相比，藏文自动问答方面的研究相对滞后。特别是供研究和应用的高质量藏文自动问答数据集匮乏问题，成为藏文自动问答领域的发展瓶颈。因此，建立一个高质量的藏文问答数据集对藏文自动问答技术发展至关重要。

从数据集构建角度考虑，构建数据集的句子类型范围越广构建效果越好。由于藏文句法结构复杂，抽取出所有句型的语法特点自动构建包含所有句型的数据集有一定的困难。选择某一特定句型抽取其句法特征构建数据集，之后将该方法推广到其他句型数据集构建，这种从特殊到一般的数据集构建方法既有针对性，同时可以从特殊句型的数据集构建方法中归纳出一般句型数据集构建的方法。

藏文句法结构中位格助词、作格助词、属格助词以及从格助词等特殊助词扮演着重要角色，才智杰教授在承担的国家社会科学基金项目“面向自然语言处理的藏文句型结构分布统计研究”(13BYY141)[4] 中对藏文句型结构进行了分类分布统计，得出仅包含一个位格助词且动词结尾的句型在藏文句型结构中出现频率最高，是藏文句子的主要句型，在藏语语法中具有代表性。因此，本文选择出现频率最高的仅包含一个位格助词且动词结尾的单句作为研究对象，并结合特殊助词在语法上不同的功能构建句子及问句模板，完成了面向自动问答的藏文动词结尾型数据集构建。为了便于描述，下文将仅包含一个位格助词且动词结尾的单句用“动词结尾+位格助词”型或“V+GL”型表示。

1　相关研究

自动问答数据集是研究问答技术的数据资源，由数据源、问题和答案 3 部分组成。截至目前，学者们已建立了许多大规模英文自动问答数据集。SQuAD[5] 的数据源是维基百科的文章，问题和答案由人工建立，包含 107 785 个问答对。MS MARCO[6] 的数据源是搜索引擎查询出的相关内容，问题是用户在搜索引擎的查询记录，答案由人工建立，包含 100 000 个问答对。TriviaQA[7] 的数据源是搜索引擎查询出的相关内容，问题和答案是问答网站已存的问答对，包含 950 000 个问答对。DuoRC[8] 的数据源是维基百科和 IMDb 网站对同一部影片的两种描述，问题由人工根据其中一个描述提出，答案由人工根据另一个描述回答，包含 186 089 个问答对。NarrativeQA[9] 的数据源是书籍和电影剧本，问题和答案由人工根据维基百科对应故事的描述建立，包含 15 670 个问答对。RACE[10] 的数据源是中国英语考试文章，问题由人工提出，答案由人工在多个候选答案中选择最相关的，包含 97 867 个问答对。MCTest[11] 的数据源是人工编写的虚拟故事，问题由人工提出，答案由人工在多个候选答案中选择最相关的，包含 2 000 个问答对。CosmosQA[12] 的数据源是博客网站的文章段落，问题由人工提出，答案由人工在多个候选答案中选择最相关的，包含 35 600 个问答对。CNN/Daily Mail[13] 的数据源是 CNN 和 Daily Mail 网站的新闻，问题是缺少一个单词或短语的句子，答案是缺少的单词或短语，包含 1 260 000 个问答对。Who-did-What[14] 的数据源是 Gigaword 语料库，问题是移除单词的句子，答案是移除的单词，包含 20 000 个问答对。英文自动问答数据集信息见表 1。

表 1　英文自动问答数据集信息

Table 1　Information about the automatic questions and answers dataset in English

数据集	构建者	数据源	问题	答案	规模
SQuAD	Pranav Rajpurkar	维基百科文章	人工提出	人工回答	107 785
MS MARCO	Payal Bajaj	搜索引擎结果	查询记录	人工回答	100 000
TriviaQA	Mandar Joshi	搜索引擎结果	现有问题	现有答案	950 000
DuoRC	Amrita Saha	电影描述	人工提出	人工回答	186 089
NarrativeQA	Tomáš Kočiský	书籍和电影剧本	人工提出	人工回答	15 670
RACE	Guokun Lai	英语考试文章	现有问题	现有答案	97 867

续表 1

数据集	构建者	数据源	问题	答案	规模
MCTest	Matthew Richardson	人工编写故事	人工提出	人工选择	2 000
CosmosQA	Lifu Huang	博客	现有问题	现有答案	35 600
CNN/Daily Mail	Karl Moritz Hermann	新闻网站新闻	删除片段	单词短语	1 260 000
Who-did-What	Takeshi Onishi	Gigaword语料	删除片段	单词短语	20 000

汉文自动问答数据集有 CMRC[15]、DRCD[16]、DuReader[17]、MATINF[18]、JEC-QA[19]、C3[20] 和 ChID[21]。其中, CMRC 和 DRCD 的数据源是中文维基百科中的段落, 问题和答案由人工建立, CMRC 包含 20 000 个问答对, DRCD 包含 30 000 个问答对; DuReader 的数据源是百度知道的文本段落, 问题来自百度搜索引擎的真实问题, 答案由人工手动生成, 包含 200 000 个问答对; MATINF 的数据源是外部知识, 问题和答案是母婴网站已存在的问答对, 包含 1 070 000 个问答对; JEC-QA 的数据源是中国司法考试的文章, 问题由人工提出, 答案由人工在多个候选答案中选择最相关的答案, 包含了 26 365 个问答对; C3 的数据源是汉文考试的文章, 问题由人工提出, 答案由人工在多个候选答案中选择最相关的答案, 包含 19 577 个问答对; ChID 的数据源是新闻和小说, 问题是从句子中删除成语的句子, 答案是句子中应该添加的成语, 包含了 729 000 个问答对。汉文自动问答数据集信息见表 2。

表 2　汉文自动问答数据集信息

Table 2　Information about the automatic question-answer dataset in Chinese

数据集	构建者	数据源	问题	答案	规模
CMRC	Cui Yiming	维基百科文章	人工提出	人工回答	20 000
DRCD	Shao Chih Chieh	维基百科文章	人工提出	人工回答	30 000
DuReader	He Wei	百度知道	现有问题	现有答案	200 000
MATINF	Xu Canwen	外部知识	现有问题	现有答案	1 070 000
JEC-QA	Zhong Haoxi	司法考试文章	现有问题	现有答案	26 365
C3	Sun Kai	汉语考试文章	现有问题	现有答案	19 577
ChID	Zheng Chujie	新闻和小说	删除成语	缺失成语	729 000

藏文自动问答数据集比较匮乏, 目前有 TibetanQA 数据集[22-23] 可供藏文自动问答研究使用。TibetanQA 数据集的数据源是云藏网的文章, 问题和答案由人工建立, 包含 20 000 个问答对, 其中 2 000 个已公开。虽然 TibetanQA 为藏文自动问答提供了基础资源, 但其并未深入研究藏文多样性的句型结构, 限制了该数据集面对特定句型时的表现效果。

从以上研究可以看到, 目前自动问答数据集基本采用人工方式构建, 而且未针对不同句型分类构建数据集。本文面向“动词结尾+位格助词”型, 通过设计句子及问句模板, 以自动生成的方式构建了含 36 414 对问答句的藏文自动问答数据集 TiQuAD_36414。面向特定句型的数据集在研究自动问答技术时更具针对性, 并且采用自动化方式构建能解决耗时费力的问题。

2　藏文自动问答数据集的构建

2.1　藏文自动问答数据集构建方案

本文首先整理了青海师范大学自然语言处理组建立的多个语料库 (涵盖百科、教材、新闻、文学和政治等领域), 对整理的语料进行分句并剔除含有非藏文字符的句子, 并利用现有的词法分析器对句子进行分词和词性标注; 其次将句子中的名词与其修饰成分组成的名词性短语分为不同的类型, 以不同类型的名词性短语建立了句子和问句模板; 最后从语料中筛选出符合句型的句子, 基于句子及问句模板以自动化方式构建藏文自动问答数据集。面向自动问答的藏文数据集构建方案如图 1 所示。

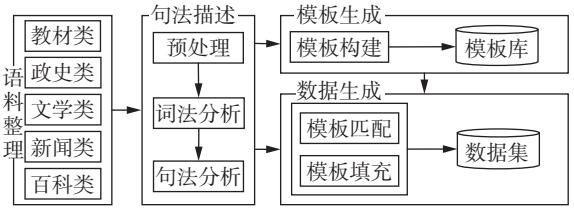

图 1　面向自动问答的藏文数据集构建方案

Fig. 1　Tibetan dataset construction scheme for Tibetan automatic question and answer

2.2　藏文自动问答数据集构建

根据藏文自动问答数据集构建方案, 对收集整理的语料进行句法描述, 从句法描述后的语料

中抽取到了 9 200 句“V+GL”型句子，通过基于模板的方法构建了面向自动问答的“V+GL”型藏文问答数据集 TiQuAD_36414。TiQuAD_36414 数据集语料类型和句型分布见表 3 和表 4。

表 3　TiQuAD_36414 数据集语料类型分布

Table 3　Distribution of corpus types in TiQuAD_36414 dataset

语料类型	数据源	问题	答案	规模
百科类	云藏百科文章	自动生成	自动生成	10 675
教材类	小学和初中课本	自动生成	自动生成	2 107
新闻类	新闻网站文章	自动生成	自动生成	22 109
文学类	经典散文和小说	自动生成	自动生成	987
政治类	江泽民文选	自动生成	自动生成	536
总数	—	—	—	36 414

表 4　TiQuAD_36414 数据集句型分布

Table 4　Distribution of sentence types in TiQuAD_36414 dataset

句型	类型说明	句子数	问题数	占比/%
ST1	动作所指向的对象	4 858	18 024	49.82
ST2	动作或行为发生的地点	1 072	4 170	11.21
ST3	动作或行为发生的时间	410	2 648	7.09
ST4	动作所要达到的目的	1 847	7 640	21.33
ST5	动作的变换结果或状态	373	1 644	4.43
ST6	动作的领有者或获得者	640	2 288	6.12
总数	—	9 200	36 414	100.00

由表 3 可见，数据集 TiQuAD_36414 涵盖了百科类、新闻类、教材类、文学类和政治类等多领域文本，其中百科类、新闻类语料占比最高，教材类、文学类和政治类语料虽然在数量上相对较少，但其对特定领域的语言特征建模具有重要的语言学价值。由表 4 可见，数据集 TiQuAD_36414 中 ST1 型句型以 49.82% 的占比居于首位，ST2 型和 ST4 型句型分别占 21.33% 和 11.21%，这 3 类句型基本覆盖了问答系统中的常见句型结构；虽然 ST3、ST5、ST6 型句型所占比例相对较小，但这些句型的引入显著提升了数据集的多样性，同时也在一定程度上增加了问答匹配任务的复杂性。

接下来介绍数据集 TiQuAD_36414 具体构建过程。

2.2.1　语料收集整理

为了构建高质量的藏文自动问答数据集，本文整理了青海师范大学自然语言处理组建立的多个语料库，涵盖百科、新闻、文学、教材和政治等领域，共计 333 772 条句子。通过对整理的语料进行剖析，得到藏文中出现频率最高的“动词结尾+位格助词”型句子的特征，进而设计构建句子与问句模板，同时从语料库中筛选符合句型结构的句子，自动构建藏文问答数据集。新闻类语料中“V+GL”型句子有 5 558 句，百科类、教材类、文学类和政治类数量共有 3 632 句，语料信息见表 5。

表 5　藏文语料信息

Table 5　Tibetan corpus information

语料类型	来源	句数	V+GL型句数
百科类	云藏百科文章	76 233	2 590
教材类	小学和初中课本	34 098	621
新闻类	新闻网站文章	134 467	5 568
文学类	经典散文和小说	41 608	276
政治类	江泽民文选	47 366	145
合计	—	333 772	9 200

2.2.2　句法描述

对于收集整理的语料做句法描述处理步骤如下。

1) 预处理

藏文以“།”或“།།”或“།+空格”或“།།+空格”作为句子的结束符号，在预处理过程中，本文首先使用单垂符“།”、双垂符“།།”和空格为断句边界点对藏文文本进行分句，然后剔除含有英文、汉字、标点符号等非藏文字符的句子。

2) 词法分析

本文采用《央金藏文分词系统》[24] 对预处理后的句子进行分词和词性标注。模板中用到的藏文词类标记见表 6。

表 6　藏文词类标记

Table 6　Tibetan part-of-speech tags

词性	标记	说明	词性	标记	说明
名词	nn	一般名词	格助词	gx	作格助词
	nr	人名		gz	属格助词
	nv	动名词		gl	位格助词
	nm	民族名		gj	从格助词
	ns	地名	数词	m	计数功能
	ng	国名	量词	q	与数词结合
	nt	机构团体	代词	rr	人称代词
	nz	专有名词		rz	指示代词
	nx	职位名		ry	疑问代词
	na	形名词	时间词	t	时间与朝代
	nl	缩名词	方位词	f	方向位置
形容词	ad	单音节	动词	vt	及物动词
	ac	多音节		vi	不及物动词
副词	d	修饰成分		ve	存在动词
助词	u	语法结构		vd	趋向动词

3) 句法分析

句法分析是构建模板的基础, 藏文语法复杂, 含有大量修饰成分, 使得同类型句子可能因修饰成分而难以成功匹配模板, 给模板构建和匹配带来巨大挑战。为了便于构建藏文中使用频率最高的动词结尾型句子模板, 本文通过分析藏文名词与其修饰成分 (形容词、副词、数词、量词、代词) 之间的关系, 将名词与这些修饰成分组成的名词性短语划分为指人短语、地点短语、时间短语和一般名词短语, 从而简化句子结构并提高模板的命中率。名词与修饰成分组成的名词性短语类型及修饰关系见表 7。其中"|"表示或, "[]"表示可选, "+"表示结合 (下同)。"NR"表示指人短语, "NS"表示地点短语, "NT"表示时间短语, "NP"表示一般名词短语。例如句子"བར་སྣང་/ffཎ་/glཁང་ཁང་/nnཙམ་མ་/qཀཱ་ཏ་/mའགྲུར་/vi"(一群大雁在空中飞翔) 的句法分析结果是"བར་སྣང་/NSཎ་/glཁང་ཁང་ཙམ་མ་ཀཱ་ཏ་/NPའགྲུར་/vi"。

表 7　名词性短语类型及修饰关系

Table 7　Noun phrase types and modifying relationships

序号	类型	修饰关系
1	NR	nr|nt|nx|rr+[([d]+<a>)|([q]+<m>)]
2	NT	t+[([d]+<a>)|([q]+<m>)]
3	NS	(ns|ng+[f])|(nn|rz+f)+[([d]+<a>)|([q]+<m>)]
4	NP	nn|nv|nm|nz|nl|na+[([d]+<a>)|([q]+<m>)]

2.2.3　句子模板构建

"V+GL"型在藏文句型结构中出现频率最高, 涵盖了藏文问答中的核心语义单元。因此, 本文以"V+GL"型构建句子模板, 生成具有一致性和扩展性的问答数据集。

虚词作为在藏文的语法结构中不可或缺的组成部分, 虽然自身不含语义, 但它在句子中起到连接词语、标志句法关系和语气等功能。藏文虚词主要包括位格助词、作格助词、属格助词以及从格助词等, 每一种都承载着特定的语法和语义功能, 其中位格助词包括"ས་ཅ་ར་ལ་ཏུ་དུ་ན"7 种, 通常添加在名词性短语后面, 表示动作的对象、地点、时间、目的和状态等。根据《实用藏文文法教程》[25] 及收集整理语料中"V+GL"型句子结构特征的分析、归纳、总结, 得到 8 种不同的语法作用。

1) ST1 型: 表示动作所指向的对象, 用位格助词"ལ་ར"之一, 位格助词前的名词性短语为指人短语或一般名词短语, 位格助词后为指人短语或一般名词短语或及物动词或不及物动词。

2) ST2 型: 表示动作或行为发生的地点, 用位格助词"ས་ཅ་ར་ལ་ཏུ་དུ་ན"之一, 位格助词前的名词性短语为地点短语, 位格助词后为指人短语或一般名词短语或及物动词或不及物动词。

3) ST3 型: 表示动作或行为发生的时间, 用位格助词"ས་ཅ་ར་ལ་ཏུ་དུ་ན"之一, 位格助词前的名词性短语为时间短语, 位格助词后为指人短语或一般名词短语或及物动词或不及物动词。

4) ST4 型: 表示动作的变换结果或动作的状态, 用位格助词"ས་ཅ་ར་ལ་ཏུ་དུ"之一, 位格助词前是一般名词短语或形容词, 位格助词后是及物动词和不及物动词。

5) ST5 型: 表示动作所要达到的目的, 用位格助词"ས་ཅ་ར་ལ་ཏུ་དུ"之一, 位格助词前是及物动词, 位格助词后是趋向动词。

6) ST6 型: 表示动作的领有者或获得者, 用位格助词"ས་ཅ་ར་ལ་ཏུ་དུ་ན"之一, 位格助词前的名词性短语为指人短语或一般名词短语, 位格助词后是一般名词短语, 句尾一般是存在动词。

7) ST7 型: 表示短语之间的关联成分, 用位格助词"ཨ",位格助词的前后都是形容词。

8) ST8 型: 表示复句中两个分句之间的关系, 用位格助词"ཨ", 位格助词出现在第一个分句的句末。

ST7 型的位格助词做两个形容词之间的关联词, 相当于汉文中的"既……又……"和"又……又……"; ST8 型的位格助词用来连接复句中的两个句子, 位格助词出现在第一个分句的句末。本文的研究对象是"V+GL"型单句, 因此在构建句子模板时未考虑 ST7 型和 ST8 型。

对以上 ST1~ST6 型藏文句子进行句法描述中的词法分析和句法分析, 构建了"V+GL"型句子和问句模板, 如表 8 所示。

表 8　"V+GL"型句子和问句模板

Table 8　"V+GL" sentence and question templates

句子类型	句子模板		问句模板	
ST1	*x*+gl+*y*+vt	*x*=NP|NR, *y*=[NP]	*F*(*x*)|*R*(*x*)+gl+*y*+vt	*x*+gl+*F*(*y*)|*R*(*y*)+vt
	y+*x*+gl+vt		*y*+*F*(*x*)|*R*(*x*)+gl+vt	*F*(*y*)|*R*(*y*)+*x*+gl+vt
	x+gl+*y*+vi	*x*=NP|NR, *y*=[NP|NR]	*F*(*x*)|*R*(*x*)+gl+*y*+vi	*x*+gl+*F*(*y*)|*R*(*y*)+vi
	y+*x*+gl+vi		*y*+*F*(*x*)|*R*(*x*)+gl+vi	*F*(*y*)|*R*(*y*)+*x*+gl+vi

续表 8

句子类型	句子模板		问句模板	
ST2	*x*+gl+*y*+vt	*x*=NS, *y*=[NP]	*F*(*x*)|*R*(*x*)+gl+*y*+vt	*x*+gl+*F*(*y*)|*R*(*y*)+vt
	y+*x*+gl+vt		*y*+*F*(*x*)|*R*(*x*)+gl+vt	*F*(*y*)|*R*(*y*)+*x*+gl+vt
	x+gl+*y*+vi	*x*=NS, *y*=[NP|NR]	*F*(*x*)|*R*(*x*)+gl+*y*+vi	*x*+gl+*F*(*y*)|*R*(*y*)+vi
	y+*x*+gl+vi		*y*+*F*(*x*)|*R*(*x*)+gl+vi	*F*(*y*)|*R*(*y*)+*x*+gl+vi
ST3	*x*+gl+*y*+vt	*x*=NT, *y*=[NP]	*F*(*x*)|*R*(*x*)+gl+*y*+vt	*x*+gl+*F*(*y*)|*R*(*y*)+vt
	y+*x*+gl+vt		*y*+*F*(*x*)|*R*(*x*)+gl+vt	*F*(*y*)|*R*(*y*)+*x*+gl+vt
	x+gl+*y*+vi	*x*=NT, *y*=[NP|NR]	*F*(*x*)|*R*(*x*)+gl+*y*+vi	*x*+gl+*F*(*y*)|*R*(*y*)+vi
	y+*x*+gl+vi		*y*+*F*(*x*)|*R*(*x*)+gl+vi	*F*(*y*)|*R*(*y*)+*x*+gl+vi
ST4	*x*+gl+*y*+vt	*x*=a, *y*=[NP]	*F*(*x*)|*R*(*x*)+gl+*y*+vt	*x*+gl+*F*(*y*)|*R*(*y*)+vt
	y+*x*+gl+vt		*y*+*F*(*x*)|*R*(*x*)+gl+vt	*F*(*y*)|*R*(*y*)+*x*+gl+vt
	x+gl+*y*+vi	*x*=a, *y*=[NP|NR]	*F*(*x*)|*R*(*x*)+gl+*y*+vi	*x*+gl+*F*(*y*)|*R*(*y*)+vi
	y+*x*+gl+vi		*y*+*F*(*x*)|*R*(*x*)+gl+vi	*F*(*y*)|*R*(*y*)+*x*+gl+vi
ST5	*x*+*y*+vt+gl+vd	*x*=[NP|NR], *y*=[NP]	*F*(*x*)|*R*(*x*)+*y*+vt+gl+vd	*x*+*F*(*y*)|*R*(*y*)+vt+gl+vd
	x+vi+gl+vd	*x*=[NP|NR]	*F*(*x*)|*R*(*x*)+vi+gl+ vd	
ST6	*x*+gl+*y*+ve	*x*=[NP|NR], *y*=NP	*F*(*x*)|*R*(*x*)+gl+*y*+ve	*x*+gl+ *F*(*y*)|*R*(*y*)+ve

2.2.4　问句模板构建

现代藏文文法将句子按语气分为陈述句、疑问句、祈使句和感叹句。班玛宝等[26]通过分析疑问句结构特征发现每个疑问句至少含一个疑问代词，本文在构建问句模板时需要用如表 9 所示的藏文疑问代词。

表 9　疑问代词

Table 9　Interrogative pronoun

标记	疑问代词	标记	疑问代词
ry1	ཅི་ཞིག	ry2	གང
ry3	གང་དུས་གང་། རམ་ཞིག	ry4	ས། སུ་ཞིག་ཐུ་ཚོ་ཐུ་རྣམས
ry5	རྟེ་འདྲ	ry6	གཅམ། གཙོད

为了便于表示问句模板中的疑问代词，本文引入了函数 $R(x)$ 和 $F(x)$。函数 $R(x)$ 表示取 ry1 至 ry5 中的某一类疑问代词，$F(x)$ 表示取疑问代词 ry6 的条件，函数定义为

$$R(x)=\begin{cases}\text{ry1}, & x=\text{NP}\\ \text{ry2}, & x=\text{NS}\\ \text{ry3}, & x=\text{NT}\\ \text{ry4}, & x=\text{NR}\\ \text{ry5}, & x=\text{a}\end{cases}\tag{1}$$

$$F(x)=\text{ry6}, x=\text{NP}|\text{NR}|\text{NS}|\text{NT},\ x\in\text{m}\tag{2}$$

在构建问句模板时，本文使用表 9 所示疑问代词对已构建的句子模板中的每一个 x 和 y 提问，并通过对问句结构归纳、总结得到了“V+GL”型问句模板，“V+GL”型问句模板如表 8 所示。

表 8 所示问句模板中的变量 x 和 y 的取值与句子模板中的取值相同，函数 $R(x)$ 和 $F(x)$ 如式 (1)、(2) 所示，用于确定问句模板中的疑问代词。例如：“*x*+gl+y+vi　*x*=NS, *y*=[NP|NR]”表示的句子模板“NS+gl+vt”对应的问句模板为“ry2+gl+vi”“ry2|ry6+gl+vi”；句子模板“NS+gl+NP+vi”对应的问句模板为“ry2+gl+NP+vi”“NS+gl+ry1+vi”“ry2|ry6+gl+NP+vi”“NS+gl +ry1|ry6 +vi”；句子模板“NS+gl+NR+vi”对应的问句模板为“ry2+gl+NR+vi”“NS+gl+ry4+vi”“ry2|ry6+gl+NR+vi”“NS+gl+ry4|ry6+vi”。

2.2.5　数据生成

在前 4 步的基础上，本文从原始语料中抽取了“V+GL”型句子，通过对抽取到的句子进行句法描述，将其表示为句子模板中的形式，根据句子模板及对应的问句模板生成问句，从而构建了面向自动问答的藏文“V+GL”型数据集 TiQuAD_36414。例如，抽取的“V+GL”型句子“བར་སྣང་དུ་ཁྱོ་ཐུང་ཚོགས་གཅིག་འཕྲུར”的词法分析结果为“བར་སྣང་/ffདུ/glཁྱོ་ཐུང་/nnཚོགས་/qཅིག་/mའཕྲུར/vi”，句法分析结果为“བར་སྣང་/NSདུ/glཁྱོ་ཐུང་ཚོགས་ཅིག་/NPའཕྲུར/vi”，句法分析结果与句子模板 (表 8) 中 ST2 型的“*x*+gl+*y*+vi　*x*=NS, *y*=[NP|NR]”模板匹配，匹配的句子模板对应的问句模板为“*F*(*x*)|*R*(*x*)+gl+*y*+vi”和“*x*+gl+*F*(*y*)|*R*(*y*)+ vi”，从而可生成问句“གང་དུ་ཁྱོ་ཐུང་ཚོགས་གཅིག་འཕྲུར”“བར་སྣང་དུ་ཅི་ཞིག་འཕྲུར”“བར་སྣང་དུ་ཁྱོ་ཐུང་གང་ཙམ་འཕྲུར”和“བར་སྣང་དུ་ཁྱོ་ཐུང་གང་ཚོད་འཕྲུར”。

2.3　TiQuAD_36414 数据集有效性分析

本文通过主观和客观的评价方法，从多个维度对面向自动问答的藏文数据集 TiQuAD_36414 的有效性进行全面分析。

2.3.1　主观评价

平均意见得分 (mean opinion score, MOS) 评价具有能够直观反映用户真实感受的优点，并且

在多人参与评价时能够保持统一的评价标准, 是一种常用的主观评价方法。本文在主观评价数据集 TiQuAD_36414 时使用 MOS 评价方法, 具体评价过程如下。

1) 建立评价项和评价标准

本文从疑问代词选择是否正确、疑问代词位置是否恰当、问题与答案是否相关和语句是否流畅等 4 个方面对生成的数据集进行有效性分析, 评价项及评价标准如表 10 所示。

表 10　评价项及评价标准

Table 10　Evaluation items and criteria

评价项	评价标准及得分		
疑问代词的选择	不正确 [0,1]	基本正确 (1,3]	正确 (3,5]
疑问代词的位置	不恰当 [0,1]	基本恰当 (1,3]	恰当 (3,5]
问题与答案的相关性	不相关 [0,1]	基本相关 (1,3]	相关 (3,5]
语句的流畅度	不流畅 [0,1]	基本流畅 (1,3]	流畅 (3,5]

一个问句的最终得分利用加权平均值计算, 权重分别为 0.2、0.2、0.2 和 0.4。这是因为疑问代词选择不正确或疑问代词位置不恰当都会降低语句的流畅度, 所以增加了语句流畅度的权值, 降低了疑问代词选择和疑问代词位置的权值。得分区间可分为“优秀”“良好”“一般”和“较差”4 个级别, (3,5] 为优秀, (2,3] 为良好, (1,2] 为一般, [0,1] 为较差, 其中“优秀”“良好”“一般”级别的问答对为合格, “较差”级别的问答对为不合格。

2) 实验数据抽样

本文从数据集 TiQuAD_36414 的 ST1~ST6 型中各随机抽取了 5% 的句子以及问句 (460 对), 作为实验评价数据。

3) MOS 评价

本文组织了 7 名从事藏语自然语言处理的研究生对抽取的 460 个问答对使用表 10 中的评价标准打分, 每个问答对的最终得分取 7 人的平均值。最终得分分布见表 11 及图 2。

表 11　问答对得分分布

Table 11　Question and answer pair score distribution

类型	[0,1]		(1,2]		(2,3]		(3,4]		(4,5]	
	句子数	占比/%	句子数	占比/%	句子数	占比/%	句子数	占比/%	句子数	占比/%
ST1	13	2.82	6	1.30	12	2.61	50	10.87	212	46.09
ST2	1	0.22	1	0.22	2	0.43	2	0.43	17	3.70
ST3	1	0.22	1	0.22	2	0.43	5	1.09	12	2.61
ST4	4	0.87	0	0.00	3	0.66	7	1.52	28	6.08
ST5	1	0.22	0	0.00	1	0.22	15	3.26	23	5.00
ST6	0	0.00	1	0.22	0	0.00	7	1.52	33	7.17
总计	20	4.35	9	1.96	20	4.35	86	18.69	325	70.65

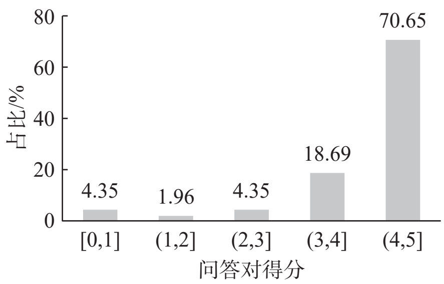

图 2　问答对得分分布

Fig. 2　Question and answer pair score distribution

表 11 和图 2 中的数据显示, 数据集中的 411 对问答句的得分在 (3,5] 内, 为“优秀”等级, 占比 89.34%; 数据集中的 20 对问答句的得分在 (2,3] 内, 为“良好”等级, 占比 4.35%; 数据集中的 9 对问答句的得分在 (1,2] 内, 为“一般”等级, 占比 1.96%; 数据集中的 20 对问答句的得分在 [0,1] 内, 为“较差”等级, 占比 4.35%。在 460 对问答句中, “优秀”“良好”和“一般”级别的总和为 440 句, 合格率为 95.65%; “较差”级别问答对的数量分别为 20, 不合格率为 4.35%, 整体而言构建的数据集合格。通过对不合格的问答对分析, 发现错误原因是词法分析不正确。例如, “ཕྱུགས་རྫིས་རྟ་ལ་སྲབ་མཐུར་བསྐོན”(牧民给马套上辔头) 词法分析结果为“ཕྱུགས་རྫི/nnས་རྟ་/mལ་/glསྲབ་མཐུར་/nnབསྐོན/vt”, 而正确结果应该为“ཕྱུགས་རྫི/nnས་/gxརྟ་/nn ལ་/glསྲབ་མཐུར་/nn བསྐོན/vt”, 从而在模板匹配时出现了错误, 这类错误今后可以通过改进词法分析器性能得以解决。本文对数据集 TiQuAD_36414 中的不合格问答对对应的句子的词法分析结果进行了人工校对, 之后生成的问答对都为“优秀”等级。

2.3.2　客观评价

在客观评价方面, 本文采用 BiDAF[27]、RNet[28] 和 QANet[29] 模型的正确率 (exact match, EM)、精

确率与召回率的调和平均数 F1 值为评价指标，以英文问答数据集 SQuAD[5] 和藏文问答数据集 TibetanQA[22-23] 上的性能作为参考值，分析数据集 TiQuAD_36414 的有效性。实验时将数据集 TiQuAD_36414 随机划分为 80% 的训练集和 20% 的测试集。实验结果和模型参数见表 12~13。

表 12　不同模型在多个数据集上的表现
Table 12　Performance of different models across multiple datasets　%

模型	SQuAD		TibetanQA		TiQuAD_36414	
	EM	F1	EM	F1	EM	F1
BiDAF	68.0	77.3	58.6	67.8	66.8	81.5
RNet	71.3	79.7	55.8	63.4	71.0	80.9
QANet	73.6	82.7	57.1	66.9	72.9	85.6

表 13　模型参数
Table 13　Model parameters

参数名称	BiDAF	RNet	QANet
词向量维度	300	300	300
学习率	0.01	0.01	0.01
训练轮数	10	10	10
批次大小	32	32	32
上下文上限	500	500	500
问句上限	50	50	50
答案上限	40	40	40
字符上限	10	10	10
优化器	Adadelta	Adam	Adadelta

表 12 显示，BiDAF、RNet 和 QANet 模型在本文构建的数据集 TiQuAD_36414 上的 EM 值分别为 66.8%、71.0% 和 72.9%，F1 值分别为 81.5%、80.9% 和 85.6%；在藏文数据集 TibetanQA 上的 EM 值分别为 58.6%、55.8% 和 57.1%，F1 值分别为 67.8%、63.4% 和 66.9%；在英文数据集 SQuAD 上的 EM 值分别为 68.0%、71.3% 和 73.6%，F1 值分别为 77.3%、79.7% 和 82.7%。以英文数据集 SQuAD 和藏文数据集 TibetanQA 上模型 BiDAF、RNet 和 QANet 的性能作为参考值可以看到，本文构建的数据集上模型 BiDAF、RNet 和 QANet 的性能都比较好，表明数据集 TiQuAD_36414 有效。

BiDAF、RNet 和 QANet 模型在数据集 TiQuAD_36414 上有多个 EM 和 F1 值高于在藏文数据集 TibetanQA 上的值，其主要原因为：1) 本文构建的数据集 TiQuAD_36414 不但针对“V+GL”型特定句型，而且规模也大。2) 本文基于模板自动从数据源中抽取句子并据此生成问题，BiDAF、RNet 和 QANet 模型也是从数据源中抽取句子，模型从数据源中抽取句子时更容易理解这一相似过程。3)TiQuAD_36414 数据集使用了基于模板的构建方法，与现有数据集相比，该数据集不仅实现了更高精度的标注一致性，还有效降低了噪声数据的比例，这使得模型在训练过程中能够获得更加准确和清晰的信息，从而提高了模型的性能表现。

3　结束语

本文在分析英文和汉文自动问答数据集构建现状的基础上，设计了“V+GL”型藏文句子模板和对应问句模板，依据句子模板和问句模板构建了“V+GL”型藏文自动问答数据集 TiQuAD_36414，并采用平均意见得分 (MOS) 方法，从疑问代词的位置、疑问代词的选择、问题与答案的相关性和语句的流畅度 4 个方面对构建的数据集进行了有效性分析，同时用 BiDAF、RNet 和 QANet 模型的 F1 值和 EM 值对数据集的有效性进行了验证。实验结果表明，构建的藏文自动问答数据集 TiQuAD_36414 性能良好。本文仅面向“V+GL”型句子构建了 36414 句对的自动问答数据集，在未来的工作中，将进一步面向其他句型构建句子和问句模板，以扩充数据集的规模，为藏文自动问答研究提供数据基础。

参考文献：

[1] 文森, 钱力, 胡懋地, 等. 基于大语言模型的问答技术研究进展综述[J]. 数据分析与知识发现, 2024, 8(6): 16−29.
WEN Sen, QIAN Li, HU Maodi, et al. Review of research progress on question-answering techniques based on large language models[J]. Data analysis and knowledge discovery, 2024, 8(6): 16−29.

[2] 王娜, 李杰. 基于 AHP-熵权法的 FAQ 问答系统用户满意度评价研究: 以高校图书馆问答型机器人为例[J]. 情报科学, 2023, 41(9): 164−172.
WANG Na, LI Jie. User satisfaction evaluation of FAQ system based on AHP-entropy weight method: taking the question answering robot of university library as an example[J]. Information science, 2023, 41(9): 164−172.

[3] 车万翔, 窦志成, 冯岩松, 等. 大模型时代的自然语言处理: 挑战、机遇与发展[J]. 中国科学: 信息科学, 2023, 53(9): 1645−1687.
CHE Wanxiang, DOU Zhicheng, FENG Yansong, et al. Towards a comprehensive understanding of the impact of large language models on natural language processing: challenges, opportunities and future directions[J]. Scientia sinica (informationis), 2023, 53(9): 1645−1687.

[4] 才智杰. 面向自然语言处理的藏文句型结构分布统计(13BYY141)研究工作报告[R]. 青海: 国家社科基金项目, 2016.

[5] RAJPURKAR P, ZHANG Jian, LOPYREV K, et al. SQuAD: 100, 000+ questions for machine comprehension of text[C]//Proceedings of the 2016 Conference on Empirical Methods in NaturalLanguage Processing. Austin: ACL, 2016: 2383−2392.

[6] BAJAJ P, CAMPOS D, CRASWELL N, et al. MS MARCO: a human generated MAchine reading COmprehension dataset[EB/OL]. (2018−10−31)[2024−10−02]. https://arxiv.org/abs/1611.09268v3.

[7] JOSHI M, CHOI E, WELD D, et al. TriviaQA: a large scale distantly supervised challenge dataset forReading comprehension[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver: ACL, 2017: 1601−1611.

[8] SAHA A, ARALIKATTE R, KHAPRA M M, et al. DuoRC: towards complex language understanding with paraphrased reading comprehension[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne: ACL, 2018: 1683−1693.

[9] KOČISKÝ T, SCHWARZ J, BLUNSOM P, et al. The NarrativeQA reading comprehension challenge[J]. Transactions of the association for computational linguistics, 2018, 6: 317−328.

[10] LAI Guokun, XIE Qizhe, LIU Hanxiao, et al. RACE: large-scale ReAding comprehension dataset from examinations[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen: ACL, 2017: 785−794.

[11] RICHARDSON M, BURGES C J C, RENSHAW E. MCTest: a challenge dataset for the open-domain machine comprehension of text[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL, 2013: 193−203.

[12] HUANG Lifu, LE BRAS R, BHAGAVATULA C, et al. Cosmos QA: machine reading comprehension with contextual commonsense reasoning[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019: 2391−2401.

[13] HERMANN K M, KOČISKÝ T, GREFENSTETTE E, et al. Teaching machines to read and comprehend[J]. Advances in neural information processing systems, 2015, 28: 1693−1701.

[14] ONISHI T, WANG Hai, BANSAL M, et al. Who did what: a large-scale person-centered cloze dataset[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: ACL, 2016: 2230−2235.

[15] CUI Yiming, LIU Ting, CHE Wanxiang, et al. A span-extraction dataset for Chinese machine reading comprehension[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: ACL, 2019: 5883−5889.

[16] SHAO C C, LIU T, LAI Yuting, et al. DRCD: a Chinese machine reading comprehension dataset[EB/OL]. (2019−05−29)[2024−10−02]. https://arxiv.org/abs/1806.00920v3.

[17] HE Wei, LIU Kai, LIU Jing, et al. DuReader: a Chinese machine reading comprehension dataset from real-world applications[C]//Proceedings of the Workshop on Machine Reading for Question Answering. Melbourne: ACL, 2018: 37−46.

[18] XU Canwen, PEI Jiaxin, WU Hongtao, et al. MATINF: a jointly labeled large-scale dataset for classification, question answering and summarization[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S. l.]: ACL, 2020: 3586−3596.

[19] ZHONG Haoxi, XIAO Chaojun, TU Cunchao, et al. JEC-QA: a legal-domain question answering dataset[J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(5): 9701−9708.

[20] SUN Kai, YU Dian, YU Dong, et al. Investigating prior knowledge for challenging Chinese machine reading comprehension[J]. Transactions of the association for computational linguistics, 2020, 8: 141−155.

[21] ZHENG Chujie, HUANG Minlie, SUN Aixin. ChID: a large-scale Chinese IDiom dataset for cloze test[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019: 778−787.

[22] 孙媛, 旦正错, 刘思思, 等. 面向机器阅读理解的藏文数据集 TibetanQA[J]. 中国科学数据, 2022, 7(2): 34−42.
SUN Yuan, DAN Zhengcuo, LIU Sisi, et al. TibetanQA: a dataset of Tibetan for Machine reading comprehension[J]. China scientific data, 2022, 7(2): 34−42.

[23] 孙媛, 刘思思, 陈超凡, 等. 面向机器阅读理解的高质量藏语数据集构建[J]. 中文信息学报, 2024, 38(3): 56−64.
SUN Yuan, LIU Sisi, CHEN Chaofan, et al. Construction of high-quality Tibetan dataset for machine reading comprehension[J]. Journal of Chinese information processing, 2024, 38(3): 56−64.

[24] 史晓东, 卢亚军. 央金藏文分词系统[J]. 中文信息学报, 2011, 25(4): 54−56.
SHI Xiaodong, LU Yajun. A Tibetan segmentation system: Yangjin[J]. Journal of Chinese information processing, 2011, 25(4): 54−56.

[25] 格桑居冕, 格桑央京. 实用藏文文法教程[M]. 成都: 四川民族出版社, 2004.

[26] 班玛宝, 才智杰, 拉玛扎西. 基于 PCFG 的藏文疑问句句法分析[J]. 中文信息学报, 2019, 33(2): 67−74.
BAN Mabao, CAI Zhijie, LA M. Tibetan interrogative sentences parsing based on PCFG[J]. Journal of Chinese information processing, 2019, 33(2): 67−74.

[27] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension[EB/OL]. (2018−06−21)[2024−10−02]. https://arxiv.org/abs/1611.01603v6.

[28] WANG Wenhui, YANG Nan, WEI Furu, et al. Gated self-matching networks for reading comprehension and question answering[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver: ACL, 2017: 189−198.

[29] YU A W, DOHAN D, LUONG M T, et al. QANet: combining local convolution with global self-attention for reading comprehension[EB/OL]. (2018−04−23)[2024−10−02]. https://arxiv.org/abs/1804.09541v1.

作者简介：

张洪溪，硕士研究生，主要研究方向为藏文信息处理、藏语自然语言处理。E-mail: 1036974179@qq.com。

才智杰，教授，博士生导师，博士。主要研究方向为藏文信息处理、藏语自然语言处理。发表学术论文 64 篇。E-mail: czjqhsd@163.com。