# 面向对话式阅读理解的高质量藏语数据集构建

达哇才仁[1,2]　朋毛才让[1,2]　孙媛[1,2,3,*]
[1]中央民族大学　信息工程学院，北京　100081
[2]国家语言资源监测与研究少数民族语言中心
[3]民族语言智能分析与安全治理教育部重点实验室
*通讯作者：孙媛
tracy.yuan.sun@gmail.com

## 摘要

对话式阅读理解作为对话式人工智能领域的重要研究方向，旨在使机器能够理解自然语言文本，并能够进行多轮对话以解答与文本相关的问题。随着生成式大模型的发展，该任务也成为评测大模型性能的重要指标之一。在此过程中，高质量数据集的构建成为该领域的关键任务。目前，相关算法模型在许多英语数据集上取得了显著进展，甚至超过了人类表现。然而，对于低资源语言，尤其是缺乏相应数据集的藏语，对话式阅读理解研究尚处于起步阶段。本文采用了一种人工与半自动结合的方法策略，构建了藏语对话式阅读理解数据集TiconvQA（Tibetan Conversational Question Answering）。该数据集共包含了20,358个对话对，涵盖了人物、地理和新闻三个领域。每一轮对话包括对话依据文本以及根据文本生成的多轮连续问答对。本文从对话数据的多样性、相关性、语言现象等方面对TiconvQA进行了详尽的分析与质量评估。并对藏文对话式阅读理解任务中存在影响评价指标的五种因素进行了优化。最终，我们采用了三种经典的对话式阅读理解模型以及藏文大模型TiLamb对数据集进行实验评估，实验结果验证了数据集的质量，并表明TiconvQA可用于模型在对话式阅读理解任务中的性能评测。

**关键词：** 藏文；对话式阅读理解；低资源语言；数据集

# Construction of high-quality Tibetan language dataset for conversational reading comprehension

**DawaCairen[1,2] PengmaoCairang[1,2] Yuan Sun[1,2,3,*]**
[1] School of Information Engineering, Minzu University of China, Beijing　100081
[2] National Language Resources Monitoring and Research Center for Minority Languages
[3] Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE
*Corresponding author: Yuan Sun
tracy.yuan.sun@gmail.com

## Abstract

Conversational Reading Comprehension, as a significant research direction in conversational artificial intelligence, aims to enable machines to comprehend natural language texts and engage in multi-turn dialogues to answer questions related to the text. With the development of generative large language models (LLMs), this task has also become a critical benchmark for evaluating model performance. In this process, constructing high-quality datasets has become a key task in this field. Currently, related algorithm models have made significant progress on many English datasets, sometimes even surpassing human performance. However, for low-resource languages, especially Tibetan, which lacks corresponding datasets, research on conversational reading comprehension is still in its early stages. This paper adopts a combined strategy of manual and

第二十三届中国计算语言学大会论文集，第314页–第325页，太原，中国，2024年7月25日至28日。
卷1：主会论文

semi-automatic methods to construct the Tibetan conversational reading comprehension dataset, TiconvQA (Tibetan Conversational Question Answering). The dataset comprises 20,358 dialogue pairs covering three domains: personalities, geography, and news. Each dialogue includes the base text and multi-turn consecutive question-answer pairs generated based on the text. We provide a detailed analysis and quality assessment of TiconvQA, focusing on the diversity, relevance, and linguistic phenomena of the dialogue data. Additionally, we optimize five factors that influence evaluation metrics in Tibetan conversational reading comprehension tasks. Finally, we experimentally evaluate three classical conversational reading comprehension models and the Tibetan LLM TiLamb on the dataset. The experimental results validate the quality of the dataset and demonstrate that TiconvQA can be used for performance evaluation of models in conversational reading comprehension tasks.

**Keywords:** Tibetan , Conversational reading comprehension , Low-resource language , Dataset

## 1 引言

机器阅读理解（Machine Reading Compre-hension,MRC）是问答任务（Question Answer,QA）最受欢迎的变体之一，在自然语言处理社区引起了很多关注 (Gupta et al., 2020)。随着近年来大型语言模型的兴起，自然语言处理领域取得了前所未有的发展，针对主流语言的各种任务，大型模型的性能已经能够与人类相媲美 (Zhao et al., 2023)。然而，对于一些特定领域和低资源语言，尤其是在跨领域或跨语言情境下，这些模型的性能仍然不尽如人意。因此，在对话式阅读理解任务中仍然存在许多待解决的困难。

传统的单轮机器阅读理解任务无法实现连贯的对话回答。当面对问题中包含指代、推理、隐喻和省略等现实交流中常见语言现象的复杂情境时，传统模型往往难以做出准确的回答。为了更好地适应真实场景的需求，更复杂的挑战随之出现，即对话式机器阅读理解（Conversational Machine Reading Comprehension, CMRC） (Chen et al., 2023)。CMRC旨在使机器能够全面理解给定的文本段落，并能够对对话中的一系列问题做出恰当的回应，尽管其任务涉及到复杂的语言理解和推理，但其本质仍然是一项数据驱动的任务。因此高质量数据集的构建是其技术发展的关键基础。近年来，大规模数据集的引入和语言模型的涌现为CMRC任务的实现提供了可能性 (Lewis et al., 2019)。然而，现有的大多数CMRC数据集主要集中在中英文领域，如多轮对话数据集CoQA (Reddy et al., 2019b)和QuAC (Choi et al., 2018)，以及百度的中文问答数据集DuReader (He et al., 2017)等。这些数据集的推出推动了CMRC领域的研究。其中，以Reddy等人发布的大规模英文机器阅读理解数据集CoQA为代表，许多学者在该数据集上提出了自己的方法和模型。根据最新的官方数据，截至2024年3月发布的模型榜单中，在该数据集上最高得分为90.7分，超过人类水平1.9分，显示出显著的优势 (Reddy et al., 2019a)。

藏语作为中国少数民族语言之一，在国内外研究藏文信息处理的科研机构中，共享的数据资源十分有限。尽管近年来随着互联网的发展，网络上出现了大量的藏文信息，并且各种藏文信息处理相关工具也得到了良好的发展 (夏天赐and 孙媛, 2018) (龙从军et al., 2015) (安波and 龙从军, 2022)，但是藏文对话式阅读理解任务仍处于起步阶段。在这一任务中，最大的难题在于缺乏公开的高质量数据集，目前公开的藏文机器阅读理解数据集只有TibetanQA (Sun et al., 2021)。因此，构建一个面向对话式阅读理解任务的高质量藏文数据集成为推动藏文CMRC任务发展的迫切需求。

本文的主要贡献如下:

(1)本文设计了一个结合人工和半自动构建数据集的方法，并成功构建了一个对话式藏语机器阅读理解数据集TiconvQA，其中包含根据2,120个藏语文本段落构建的20,358个对话问答回合，内容涵盖了地理、人物和新闻等三个不同领域。数据集对话中的每一轮包括了一个问题和答案以及答案的相关证据文本。在构建过程中我们从文章收集、问题构建和校对审

核等方面均采用严格的流程，这包括确保对话轮之间存在关联性、对话自然性以及涵盖领域内容的多样性等特征。为了满足该领域相关研究人员的数据需求，我们已经将部分数据公开在https://github.com/TBNLP/TiconvQA。

(2)本文对数据集TiconvQA进行了详尽的分析比较，包括与经典数据集的段落长度、问题长度和对话轮数的平均值对比；数据集中问答类型的分布统计；以及通过抽样计算数据集中问答过程中各类语言现象的占比分布等多个角度进行了数据集质量的验证。此外，我们在实验过程中发现了评估该任务时存在的多个影响因素，包括答案的多样性、答案长度差异、藏语特有的书写特点以及分词不一致等方面。针对这些问题，我们提出了适用于该任务及其他藏语生成式任务评估的考量方法，并优化了评价指标的计算过程，以便为后续数据集评估提供参考。

(3)本文以三种经典的对话式阅读理解模型和藏语大模型Tilamb (Zhuang et al., 2024)作为TiconvQA数据集的基线模型，并展开了实验。同时我们进行了基于对话问答历史长度的消融实验观察对话问答历史长度对模型表现的影响，成功验证了数据集质量。最终，在经典的对话式阅读理解模型上的最高F1值达到了66.5%，藏语大模型Tilamb上的表现达到了72.8%。

## 2 相关工作

大规模数据集的发布是NLP（Natural Language Processing，NLP）领域近期快速发展的关键因素之一。这些数据集不仅为模型训练提供了必要的资源，而且推动了方法的创新和进步 (Torfi et al., 2020)。数据集的质量和规模直接影响到模型的理解能力和表现。对话式阅读理解研究的激增同样归功于大规模多轮对话数据集的出现。其中最为经典的包括英文数据集CoQA和QuAC，以及中文领域的DuReader数据集等。

CoQA数据集构建于涵盖7个不同领域的文本段落之上，总计包含8,000个对话，共计126,000个问答对。每个领域的数据均来源于特定的公开数据集，例如新闻领域的文章选自CNN/Daily Mail新闻数据集 (Hermann et al., 2015)。值得注意的是，Reddit文章和科学文章这两个领域的数据仅用于模型评估，并未参与模型训练，以此构成域外评估场景；而其余五个领域的数据则同时用于模型训练和评估，形成域内评估场景。CoQA数据集中的对话模拟了两个参与者之间的问答交互过程，其中一方提问，另一方根据文本内容进行回答。问题的形式不受限制，但要求答案能够在文本中找到充分的依据并体现一定的推理能力。虽然答案同样采用自由形式，并在文本中标注了相应的支持片段，但Yatskar等人 (Choi et al., 2018)发现，答案往往只是支持片段的轻微改写版本。CoQA数据集中的对话主要集中于细节信息的深入挖掘（约占所有问题的60%），而较少涉及诸如话题转换、澄清或定义等其他对话功能。评估指标则采用了宏观平均F1分数，并分别针对域内和域外场景进行计算，以全面衡量模型的性能。

QuAC数据集包含来自14,000个信息搜索对话框的100,000个问题。对话是根据维基百科文章中关于不同个体的部分准备的。数据集采用了不对称的设置：学生只能看到文章的标题和摘要，而老师可以看到整篇文章。这种设置迫使学生根据对话中获得的有限信息来探索隐藏问题的答案，而老师则通过提供文章的简短摘录（或在无法回答时回复“不知道”）来回答问题。由于数据集的不对称性，学生无法直接从文章中复制答案，这使得问题更具描述性、高度依赖上下文且开放性。与CoQA数据集相比，QuAC中的对话更频繁地切换到新的主题，从而增加了任务的复杂性。在该数据集中的答案是抽取式的，包括“是/否”、“不知道”以及从文章中提取的文本片段。此外，QuAC数据集不仅仅提供从文章中提取的答案片段，还包含额外的对话行为标注，如继续（跟进，可能跟进，或不跟进）。这些标注信号提供了额外有用的对话流信息，可用于训练当中，使模型能够更深入地理解对话的动态 (Zaib et al., 2022)。在评估过程中，除了整个数据集的宏观平均F1分数外，QuAC还评估了人类等效分数（HEQ） (Qiu et al., 2021)，衡量系统相对于普通人类的表现，通过比较系统的F1分数与人类的F1分数来确定系统是否达到或超过人类水平。

DuReader数据集是目前最大的中文机器阅读理解数据集，与前两者不同，中文的DuReader数据集则是一个大规模、面向真实应用、由人类生成的中文阅读理解数据集。其中所有的问题、原文都来源于实际数据（百度搜索引擎数据和百度知道问答社区），答案是由人类回答的。共包含了20万个问题、100万个文档和超过42万个人类总结的答案 (He et al., 2017)。

目前，英文和中文的对话式阅读理解数据集已经取得了显著进展，然而，在低资源语言领域的研究相对较少，这严重阻碍了低资源语言的对话式阅读理解的发展。在藏语领域，目前公

开的藏文机器阅读理解数据集仅有TibetanQA。TibetanQA包含了由人工手动根据1513篇文章构建的20000个问答对，形式为单轮机制，该数据集涵盖了自然、文化和教育等多个领域的文章 (Sun et al., 2021)。TibetanQA数据集的出现使得藏语机器阅读理解领域的尝试有了基础的数据支撑，然而，在藏语对话式阅读理解领域，数据集的空缺仍然是该领域发展的首要难点。为了解决这个问题，我们构建了首个藏语对话式阅读理解数据集，即TiconvQA。

## 3 构建过程

### 3.1 数据来源

我们从三个不同的领域中选择了文本段落作为TiconvQA数据来源：分别为来自云藏百科（https://baike.yongzin.com）的人物和地理词条信息，以及来自中国藏族网通（https://ti.tibet3.com）的新闻文章。在数据的清洗和整理工作中，我们对获取的原始数据进行了降噪处理，去除了图片、表格等非文本数据，最终得到了2427篇文章。然而，并非所有这些领域的文章都同样适合引发有趣的对话，因为只有一个实体的文章通常会导致问题完全集中在那个实体上。为了确保数据的质量和多样性，我们选取的文本长度被控制在200至600个音节之间。经过严格筛选，最终确定了2120条文本作为故事素材，用于构建对话问答。这些文本涵盖了丰富的主题和情节，有助于提高对话模型的应答能力和适应性。

### 3.2 构建方法

#### 3.2.1 人工构建

为了保证数据集的质量，我们选择组内三位研究机器阅读理解方向的藏族学生进行此数据集的构建工作。对于数据集的构建规则，我们将英文领域经典的多轮对话数据集CoQA作为主要基准，参考其构建思想和方案，保证构建过程的合理性和有效性。

| | |
|---|---|
| 标题 | སྒོག་དཔོན་ཙོན་མེ་གཙུ་རྫི་ན（吞弥 • 桑布扎） |
| 段落 | སྒོག་དཔོན་ཙོན་མེ་གཙུ་རྫི་ན་ནི་བོད་ཀྱི་ཡི་གེའི་སྐོལ་ཙོག་མཁར་གཤོད་པ་པོ་ཡིན། ཁོང་ནི་སྤྱི་ལོ་དྲུག་རབས་བདུན་པའི་ཤ་ག་ཡོད་ལྷག་ར་ཁ་བར་སློན་པོ་ཙོན་མེ་ཨ་ཆའི་ཐས་གྲུབ་ཐ…… 吞弥 • 桑布扎是藏文的创制者，他出生在公元七世纪的 617 年，吞鲁热卡地方大臣吞弥 • 阿奴的家中。…… |
| 对话历史 | Q1：བོད་ཀྱི་ཡི་གེའི་སྐོལ་གཤོད་མཁན་ནི་སུ་ཡིན། 藏文的创造者是谁？<br>A1：སྒོག་དཔོན་ཙོན་མེ་གཙུ་རྫི་ན། 吞弥 • 桑布扎<br>R1：སྒོག་དཔོན་ཙོན་མེ་གཙུ་རྫི་ན་ནི་བོད་ཀྱི་ཡི་གེའི་སྐོལ་ཙོག་མཁར་གཤོད་པ་པོ་ཡིན། 吞弥 • 桑布扎是藏文的创制者<br>Q2：ཁོང་ནམ་དུས་སྐྱེ་འཁྲུངས་བ་རེད། 他何时出生？<br>A2：སྤྱི་ལོའི་དྲུག་རབས་བདུན་པའི་ཤ་ག་ཡོད་公元七世纪的 617 年<br>R2：ཁོང་ནི་སྤྱི་ལོའི་དྲུག་རབས་བདུན་པའི་ཤ་ག་ཡོད་ 在公元七世纪的 617 年<br>Q3：ཡབ་ནི་སུ་ཡིན། 是谁的儿子？<br>A3：སློན་པོ་ཙོན་མེ་ཨ་ཆ། 吞弥 • 阿奴<br>R3：སློན་པོ་ཙོན་མེ་ཨ་ཆའི་སྐས་སུ་འཁྲུངས། 大臣吞弥 • 阿奴的家中<br>…… |

图 1: 数据集构建示例

在人工构建多轮对话的过程中，由提问者和回答者两人协助完成。提问者根据文本提出问题，而回答者给出问题的回答以及答案证据。构建的数据集的示例如图1所示，其中展示了每个段落下的多轮问题与答案，展示了在多轮对话过程中，存在的指代和省略等多样的关联现象。此外，数据集中还包含了每个答案在原文中的出处，即证据。为确保数据集的质量，每个段落

第二十三届中国计算语言学大会论文集，第314页–第325页，太原，中国，2024年7月25日至28日。
卷1：主会论文

都构建了包含5-15轮问答对的对话。问题的类型涵盖了是非题、无法回答的问题以及直接从文本段落中获取答案的问题等。

### 3.2.2 半自动扩充

为了消除在数据集构建过程中的人类的不一致性和主观性，以及解决人工构建效率缓慢等问题，我们采用一种半自动的方式实现数据集的扩充。利用了大型语言模型在内容生成方面的先进能力，设计了一个基于ChatGPT3.5 (Achiam et al., 2023)的半自动数据集构建流程。随后，我们进行了人工的细致审核和校对，以确保数据集的质量和准确性。具体的构建流程如图2所示。

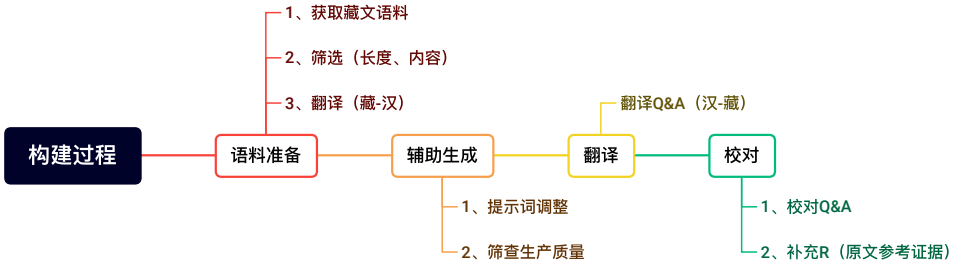


图 2: 半自动扩充构建流程

（1）预处理：首先，本文使用翻译工具将已准备好的文本转换为中文，这是因为当前最先进的大型模型主要集中在中英文等高资源语言上的研究，对于低资源语言（如藏文）的理解能力相对较弱。在这一阶段，关键在于不断调整和优化提示词，以确保生成的内容符合我们的质量标准和数据格式要求。

（2）自动化生成：完成预处理阶段后，调用ChatGPT3.5的API批量自动构建数据集。这一步骤充分利用了该模型在文本生成方面的优势，从而提高了数据集构建的效率和质量。

（3）人工审核和校对：在自动化生成多轮问答对之后，本研究通过人工进行细致的审核和校对，以确保数据集的准确性和可靠性。这一环节对于保证数据集质量至关重要，它有助于发现并纠正自动化生成过程中可能出现的错误和不一致之处。

在先前人工构建数据集的基础上，借助积累的构建经验，本文成功实施了半自动数据集构建流程，从而成功扩展了数据集规模。这种结合半自动和人工审核的方法不仅提高了数据集构建的效率，而且保证了数据集的质量。

## 3.3 构建难点与质量保证

在上述构建过程中伴随着许多挑战和难点。为了提升构建效率，同时注重数据集的质量，我们采取了一系列严格的方法和策略。以下是对这些挑战的介绍以及我们的应对措施：

（1）在人工构建过程中，常常存在主观臆断的情况，可能导致构建的数据集中的对话缺乏真实问答之间的连贯性。为了克服这一挑战，我们放弃了自问自答的构建方式，选择更加自然对答的方式，并在构建过程中随机地改变提问者和回答者的身份，通过这些方式以确保对话的构建是基于真实的交流模式和场景，尽可能模拟真实对话中的语言和逻辑交互。

（2）在人工反复构建数据集的过程中，存在陷入问答模式的风险。为了避免这种情况，我们设计了结合大型模型实现半自动扩充的方式，使得问答的过程更加自然随意。在此基础上，对话数据集的筛选和修订阶段经过严格的质量控制，以确保对话内容的合理性和准确性。通过这些措施，可以有效提高数据集构建效率和质量。

（3）数据集的校对和审核是非常关键。在完成结合半自动生成后，我们将最终结果翻译回藏文，并进行最后一轮的人工校对和质量审核。数据集最终的组成仍是以最原始的藏文文本及经过人工校对的对话内容。因此，需要对诸如翻译欠佳的问题进行修改、检查语法是否有问题、疑问词是否合理、表述是否自然流畅等质量把控。这种结合先进的自然语言处理技术和人工校对审核的方法，极大地提高了数据集构建的效率和质量。此外，这一流程还为未来其他低资源语言的数据集构建提供了一个可行的参考。

接下来，本文将着重介绍TiconvQA数据集的构建进展，并从数据集规模、问答类型、语言现象等多个角度分析我们的数据集，展示其在复杂性、多样性和对话相关性等方面的优势。

第二十三届中国计算语言学大会论文集，第314页-第325页，太原，中国，2024年7月25日至28日。
卷1：主会论文

# 4 数据分析

## 4.1 数据集规模

经过人工和半自动结合的构建策略，我们构建并扩充了TiconvQA数据集。最终，该数据集共包含来自2120篇文章的20358轮对话，具体构建情况如表1所示。

| 主题 | 构建方法 | 文章段落数 | QA对 |
|---|---|---|---|
| 人物 | 人工 | 669 | 5886 |
| 地理 | | 441 | 4304 |
| 新闻 | 半自动 | 1010 | 10168 |
| TiconvQA-总 | | 2120 | 20358 |

表 1: TiconvQA数据集及扩充部分构建情况

TiconvQA数据集的文本段落、问题和答案分别的平均字符长度为198、13.2和6.4，每篇文本的平均对话轮数为9.6轮。如表2所示，我们通过对比阅读理解任务的经典数据集SQuAD (Rajpurkar et al., 2016)，以及CoQA和QuAC发现，TiconvQA中的文本段落长度介于SQuAD和CoQA数据集之间，文本段落的长度直接影响着对话问答的轮数和难度。我们没有选择更长的文本段落，这是出于构建效率的考虑。考虑到团队规模的限制，过长的文本段落会明显降低数据集构建的速度。因此，文本的长度也间接影响了段落的平均对话轮数，TiconvQA相较于QuAC多出了约2.4轮。

| | TiconvQA | SQuAD | CoQA | QuAC |
|---|---|---|---|---|
| 段落平均长度 | 198 | 117 | 271 | 401 |
| 问题平均长度 | 13.2 | 10.1 | 5.5 | 6.5 |
| 答案平均长度 | 6.4 | 3.2 | 2.7 | 14.6 |
| 平均轮数 | 9.6 | - | 15.2 | 7.2 |

表 2: TiconvQA同经典数据集的对比

在对话式阅读理解任务中，对话分散在多个回合中，问题的平均长度侧面反映了对话中问答的关联性。在对话过程中，会存在许多省略、指代和非疑问句的连续问答，例如问题可以仅由一两个单词组成（谁？、何时？、为什么？），因此我们期望对话问题和答案比独立交互中的要短，从表中可以看出TiconvQA数据集的问题平均长度略高于CoQA和QuAC，答案平均长度则介于他们之间，在后续的构建中，我们将继续增加对话中问答的关联性。通过以上深入比较，能够初步的保证我们数据集的质量。

## 4.2 对话问答类型的分布

| 问答类型 | TiconvQA | SQuAD | CoQA |
|---|---|---|---|
| 可回答 | 93.1% | 66.7% | 98.7% |
| 不可回答 | 6.9% | 33.3% | 1.3% |
| 什么、哪个（名词短语） | 17.4% | 25.0% | 19.6% |
| 谁（命名实体） | 11.5% | 35.9% | 28.7% |
| 是否 | 9.7% | 0.1% | 19.8% |
| 多少（数字） | 23.9% | 16.5% | 9.8% |
| 何时（时间） | 18.4% | 7.1% | 3.9% |
| 其他 | 19.1% | 15.6% | 18.1% |

表 3: 问答类型统计对比

第二十三届中国计算语言学大会论文集，第314页-第325页，太原，中国，2024年7月25日至28日。
卷1：主会论文

在藏语中，每个问题类型都有着多样的复合形式。因此，在数据集构建的过程中，我们鼓励在提出问题时选择同义的其他复合疑问代词。此外，我们还模拟了实际对话中可能出现的无法回答问题和是非疑问句，以提高数据集的真实性和实用性。本文根据疑问词对提出的问题进行了统计分类。最终数据集中各问答类型的统计情况及与英文经典数据集的对比结果见表3所示。从各问答类型的分布情况来看，我们的数据集与经典数据集呈现出相似的趋势，各问答类型的比例相对均衡。值得注意的是，不可回答类型的比例高于SQuAD和CoQA数据集。

### 4.3 语言现象分析

在对话式阅读理解数据集的构建中，确保对话与文本段落以及对话历史之间的关联性，以及对话整体的连贯性至关重要。因此，我们进一步分析了对话与段落以及对话历史之间的关系。我们从数据集中抽样了150个对话问答对，并按照表4所示进行了各种现象的分析统计。

| 语言现象 | 示例 | 占比 |
|---|---|---|
| | 对话问题与段落之间的关系 | |
| 词汇匹配 | Q：ཁོ་སྨོན་ཆེན་གང་ནས་**མཐར་ཕྱིན་**པ་རེད། 他**毕业于**哪所大学？ | |
| | A：བོད་ཨྰངས་སྨོན་ཆེན། 西藏大学 | 29.33% |
| | R：བཀྲ་ཤིས་ནི་བོད་ཨྰངས་སྨོན་ཆེན་གྱི་བོད་རིག་པའི་སྡེ་ཚན་ནས་**མཐར་ཕྱིན།**<br>扎西**毕业于**西藏大学藏学系 | |
| 释义重述 | Q：ཁོ་ལ་**བརྩམས་ཚོས་**གང་དག་ཡོད། 他的**作品**有？ | |
| | A：ཤར་ཚའུ་རྒྱབ་རྒྱུད། 《青春瀑布》 | 49.33% |
| | R：ཁང་གི་**སྐྱབ་ངག་**ཤར་ཚའུ་རྒྱབ་ཀྱི་རང་བཟོས་སྐྱབ་ངག་གི་ཐོག་མ་རེད། <br>他的**诗歌**《青春瀑布》是第一首藏文自由诗 | |
| 语用学 | Q：ཆོ་རིང་ནི་བུ་ཡིན་ནམ་བུ་མོ་ཡིན། 才仁是男生还是女生？ | |
| | A：མོ། **女** | 21.34% |
| | R：**ཁོ་མོ་**མཚོ་སྔོན་ཞིང་ཆེན་ཡུལ་ཕྱལ་སོང་ཁྲིར་ནས་སྐྱེས། **她**出生在青海省玉树市 | |
| | 问题与其对话历史之间的关系 | |
| 无指代 | Q：**དགེ་ལུགས་པའི་**སྐོལ་འབྱེད་མཁན་སུ་ཡིན། **格鲁派**的创教人是？ | 22% |
| 显性指代 | Q：གཅན་ཆེན་ཐང་ལྷ་རི་རྒྱུད་ཀྱི་རི་རྩེ་གཙོ་བོ་ནི་གང་ཡིན། <br>念青唐古拉山脉的主峰是？ | 48.67% |
| | A：རི་བོ་གཅན་ཆེན་ཐང་ལྷ། 念青唐古拉山 | |
| | Q：**དེའི་**མཐོ་ཚད་ཀ་ཚོད་ཡོད། **它**有多高？ | |
| 隐性指代 | Q：གསར་འགྱུར་ཁྲབ་བསྒྲགས་ཚོགས་འདུ་ནམ་དུས་འཚོགས་རྒྱུ་རེད། <br>新闻发布会何时召开？ | 29.33% |
| | A：ས་གནས་དེ་གའི་དུས་ཚོད་ཕྱི་ དྲོའི་ཆུ་ཚོད་༣་དུས། 当地时间下午 3 点 | |
| | Q：**གང་དུ། 在哪里？** | |

表 4: 对话问题与段落之间的关系

在对话中问题与段落的关系中，我们将问题分类为词汇匹配，如果问题中至少包含一个在原文中出现的实词。这类问题约占总数的29.33%。若问题没有词汇匹配，但与原文释义相近，则归类为释义重述。这类问题涉及同义词、反义词、上下义和否定等现象，约占总数的49.33%。其余21.34%的问题没有词汇线索，我们将其分类为语用学，这些包括常识和预设等

现象。例如表中的证据“她出生在青海省玉树市”并不与问题“才仁是男生还是女生？”存在直接相关，但是结合世界知识可以回答这个问题。对于问题与其对话历史的关系，我们将问题分类为是否依赖于对话历史。如果依赖于对话历史，则检查问题是否包含明确的标记。我们的分析表明，约22%的问题不依赖于与对话历史的共指，可以单独回答。将近一半的问题48.67%包含明确的共指标记，如他、她、那个等。这些指代对话中引入的实体或事件。其余的29.33%没有明确的指代标记，但隐含地指代了一个实体或事件。

## 5 藏文对话式阅读理解任务的评价指标优化

对话式阅读理解作为评估大型模型能力的重要任务之一，与其他NLP任务的不同之处在于其注重对给定文本的全面理解，并据此做出合理的回应。相比之下，情感分析下、命名实体识别和文本分类等任务更关注特定信息的提取和分类。因此，对话式阅读理解任务的评价过程应侧重于模型对文本理解和语境处理的能力，以及问题回答的准确性、完整性和语境相关性。通过观察对比实验中预测的答案和标准答案，我们发现在计算评价指标之前，存在着影响实验结果的五种因素，分别为答案的多样性、答案富裕或简洁的长度差异、藏语特有的书写特点以及分词不一致等。这些因素的具体实例如表5所示。为此，我们针对每一项进行了处理和改进，以优化后续的评价指标，从而减少评估过程中可能带来的误差，并全面评估对话式藏文机器阅读理解模型的性能。

| 影响因素 | 预测答案 | 标准答案 |
|---|---|---|
| 答案多样性 | ཏི17**སྣེ་ལེན་བྱས།**<br>**接待了** 17 万 | **མི་ཞེང་ས**ཏི17<br>17 万**人次** |
| 答案富裕 | **དཔྱིམས་སྣ་ཚོགས་འགྲུང་ཞིང་**ཐྱི་ངོས་འདབ་འབྲུར་མེ་སྐྱོམས་པ<br>**形状多样且**表面凹凸不平 | ཐྱི་ངོས་འདབ་འབྲུར་མེ་སྐྱོམས་པ<br>表面凹凸不平 |
| 答案简洁 | ལུང་རྟོགས་ཡོན་ཏན་འབྱམ<br>隆多元丁邦 | **སློབ་དཔོན**་ལུང་རྟོགས་ཡོན་ཏན་འབྱམ<br>隆多元丁邦**老师** |
| 书写特点 | ་གཞུང་ལུགས་དངོས་ལུགས་རིག་ར་<br>理论物理学 | གཞུང་ལུགས་དངོས་ལུགས་རིག་པ།<br>理论物理学。 |
| 分词不一致 | རྒྱལ་ཁབ་ །ཚན་རིག་ །ལག་རྩལ་ །ཡར་རྒྱས་ །བྱ་དགའ་<br>国家丨科学丨技术丨进步丨奖 | རྒྱལ་ཁབ་ །ཚན་རིག་ལག་རྩལ་ །ཡར་རྒྱས་བྱ་དགའ་<br>国家丨科学技术丨进步奖 |

表 5: 影响评价指标的因素与实例

（1）基于音节点的重切分：在本研究中，我们采用了藏语分词工具Tip-Las (李亚超et al., 2015)来进行数据的预处理。Tip-Las是一种基于条件随机场模型实现了基于音节标注的开源藏文分词系统，能够满足我们在实验中对藏语分词的需求。在最终进行评估指标计算的过程中，我们发现，即使答案是直接从原文中抽取的，当答案和原文分别进行分词时，仍会出现分词不一致的情况。这是因为分词工具会根据不同长度和内容的特征进行分词，导致不同的分词结果所致。因此，当我们在原文中预测答案的范围时，预测的答案可能与标准答案的分词情况不一致。如表中的例子所示，即使正确的答案也可能因分词不一致而导致评价指标计算出现误差，影响对模型性能的准确评估。为了解决这一问题，在最终的评估阶段之前，我们选择对句子进行重切分，将原有的分词结果进行合并。借助藏文中音节点的特性，我们对预测答案和标准答案都进行基于音节点的重切分，以消除分词不一致带来的影响，并在随后的评价指标计算中进行使用。

（2）答案的统一格式处理：藏语作为一种拼音语言，其最小单位为音节。通常情况下，音节之间通过特定标记音节点来进行分割，而单或双垂符则用于确定句子的停顿或结束。在中英文的评测过程中，常规做法是剔除符号和停用词等内容后进行计算。然而，在针对藏语的评估过程中，除了去除中英文的标点符号和停用词外，我们发现了一些其他情况可能会影响评价指

第二十三届中国计算语言学大会论文集，第314页-第325页，太原，中国，2024年7月25日至28日。
卷1：主会论文

标的计算。具体来说，在直接进行评价指标计算时，我们注意到音节点同时出现在词两端、音节点出现的位置不一致或缺失，以及其他藏语符号的出现等问题。如表5中的书写特点一栏所示。然而，这些情况不会影响答案的准确性，但会对评价指标的计算产生影响。为解决这类问题，我们选择对待评估的标准答案和预测答案进行统一的格式处理。通过基于音节点的重切分后，我们可以在评估计算过程中选择剔除所有音节点以及其他藏语符号和停用词，以减少对于评价指标计算的影响。

（3）EM指标的优化：在对话式阅读理解任务中，答案并非只有一个标准形式。在测试过程中，我们观察到模型预测的答案与标准答案可能在内容上存在富裕、简洁或表述不一致但任然合适的情况，如表5所示的实例。然而，传统的评价指标未能全面考虑到这种情况，当答案不完全匹配时，如EM评分会直接归零。这种过于绝对的评价指标无法有效评估模型的性能和数据集的质量，因此需要采用更为细致的评价方法，结合从语义层面考虑答案的相似性和正确性。为解决这一问题，我们对EM评价指标进行了优化称为$\mathrm{EM}_{semantic}$。该指标选择采用基于词向量的字符串相似度方法来衡量预测答案与标准答案之间的相似程度，并评估其指标。旨在更准确地衡量答案之间的语义相似性，而不仅仅局限于字符级别的匹配。

## 6 实验

对话式阅读理解数据集的质量直接关系到模型的理解能力。为了评估我们构建的多轮问答数据集TiconvQA的质量，本文分别采用了seq2seq (Gehring et al., 2017)、PGNet (Gu et al., 2016) 和DrQA (Chen et al., 2017)三种经典的机器阅读理解模型和藏语大模型TiLamb (Zhuang et al., 2024)上进行测试。其任务的核心是给定一段文本p，对话历史记录$\{ q_1, a_1, \ldots, q_{i-1}, a_{i-1}\}$和一个问题$q_i$，任务是预测答案$a_i$。

### 6.1 实验设置

本文将这四种模型作为基线方法，将数据集按照8∶2的比例随机划分为训练集和测试集,Tilamb模型中我们将训练集作为模型的任务微调数据。最后将采用宏平均的精确率、召回率和F1得分作为实验的评估指标。$\mathrm{EM}_{semantic}$值是针对该任务优化的EM指标，用于衡量模型是否在语义层面与参考答案相匹配，而F1值则综合考虑了精确率和召回率，用于度量模型生成答案与参考答案之间的相似度。举例来说，假设有n个对话问题，如果模型能够正确回答m*个对话问题（其中回答的准确性取决语义相似性的判断，而非传统的字符级别的完全匹配）因此，可以利用公式(1)来计算$\mathrm{EM}_{semantic}$值。

$$EM_{semantic} = \frac{m^*}{n} \tag{1}$$

F1值是准确率(precision)和召回率(recall)的调和平均，准确率，召回率和F1值的计算如公式(2)-(4)所示:

$$precision = \frac{N(TP)}{N(TP) + N(FP)} \tag{2}$$

$$recall = \frac{N(TP)}{N(TP) + N(FN)} \tag{3}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

其中，N（TP）表示预测的当前答案和标准答案之间相同的词数，N（FP）表示不在标准答案中而在预测答案中的词数，N（FN）是标准答案中的词而不是预测答案中的词数。

### 6.2 数据集在不同模型上的表现

Seq2Seq模型采用编码器-解码器结构，将输入序列编码为向量，再解码为目标序列。在对话式阅读理解中，它接受段落、对话历史和问题，生成回答序列。PGNet是Seq2Seq的扩展，添加了复制机制，可直接复制输入文本中的单词，提高了处理实体名词等未出现在训练数据中的情况的能力。DrQA基于阅读理解，包括阅读器和检索器，前者理解文本提供答案，后者从文档中找到可能包含答案的候选文档。在对话式阅读理解中，DrQA能有效处理段落和对话历

史，并回答问题。TiLamb不同于前三者，是一种基于开源基座模型LLaMA2-7B上进行增量预训练得到的藏文大语言模型。

| Model | TiconvQA | | CoQA | |
|---|---|---|---|---|
| | F1 (%) | $EM_{semantic}$ (%) | F1 (%) | EM (%) |
| 人类表现 | 89.5 | 80.2 | 88.8 | - |
| seq2seq | 23.9 | 13.0 | 20.9 | 17.7 |
| PGNet | 40.5 | 30.7 | 45.2 | 38.0 |
| DrQA | **66.5** | **44.5** | 55.6 | 46.2 |
| Tilamb | **72.8** | **45.3** | - | - |

表 6: TiconvQA在模型上的表现

我们将针对藏语的评价指标优化应用于实际评估中。最终，我们得到了四种模型在对话式阅读理解任务中的表现结果，见表6所示，我们观察到TiconvQA数据集的表现与CoQA数据集相近。特别是在DrQA模型上，其F1值甚至超过了CoQA数据集的表现，达到了66.5%。然而，在EM值方面却低于CoQA数据集。这一现象的主要原因在于，我们的数据集中的段落文本长度以及对话问答的轮数都低于CoQA数据集。因此，数据集的难度相对较低，还有待进一步提高。在Tilamb上的表现为最高，达到了72.8%。通过这次实验，我们更全面地了解了所构建的多轮问答数据集TiconvQA的质量，以及不同模型在此任务上的性能和优劣。这为未来的研究提供了重要参考，可以针对数据集难度的提高和模型性能的优化进行进一步的探索和改进。

### 6.3 基于对话问答历史长度的消融实验

不同对话历史长度的输入对结果的影响能够侧面反映对话式数据集的质量。为了进一步证明我们数据集的质量，我们进行了不同长度的对话问答历史信息与当前问题进行拼接的实验。我们分别考虑了将当前问题之前h轮的对话问答历史信息拼接到输入中的情况，其中h的取值包括h=0、h=2、h=4以及h=full四种情况。这里的h=0和h=full分别表示不拼接对话问答历史信息和将当前问题之前的所有对话问答历史信息都进行拼接。最终，我们得到了以下的实验结果。

| 对话历史长度 | Seq2seq | PGNet | DrQA |
|---|---|---|---|
| 0 | 21.92% | 35.95% | 61.51% |
| 2 | **23.92**% | 40.46% | **66.49**% |
| 4 | 21.65% | **41.49**% | 65.56% |
| all | 21.26% | 37.91% | 64.76% |

表 7: 不同对话问答历史长度对F1值的影响

采用不同长度的对话问答历史输入进行模型训练和预测时，模型的性能会发生变化。如表7所示，我们发现在所有模型上，这种变化的规律十分一致：当历史对话长度为0时，性能最差；当历史对话长度为2时，性能最佳；然而，当加入更多对话历史时，并不会提升模型性能，反而可能降低其性能。这是因为加入更多的对话历史可能会引入更多与问题无关的信息，从而影响模型的性能。通过这一实验结果，进一步验证了我们数据集的特性，包括问答的难度、对话间的相关性、对话的连贯性等方面，以及其高质量的体现。

## 7 总结与展望

本文以人工和半自动相结合的方式构建了藏文对话式阅读理解数据集TiconvQA，为藏文该领域的研究提供了数据基础。对该任务的评估指标进行分析与优化，目前，TiconvQA的基线模型达到了66.5%的F1值，在Tilamb上的表现达到了77.8%的F1值，进一步缩小了与人类表现的差距。同时，TiconvQA也为藏语大型模型提供了评估其对话式阅读理解能力的数据。未来，我们将持续扩展数据集的规模，并鼓励更多的研究者参与到新模型的探索中。通过不断改进模型

和优化扩展数据集，我们可以推动低资源语言对话式阅读理解技术的发展，并为解决相关问题提供更加有效的解决方案。

## 致谢

本论文得到了国家社科基金(22&ZD035)和国家自然科学基金项目（61972436），以及中央民族大学项目(GRSCP202316, 2023QNYL22, 2024GJYY43)的资助。

## 参考文献

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Nuo Chen, Hongguang Li, Junqing He, Yinan Bao, Xinshi Lin, Qi Yang, Jianfeng Liu, Ruyi Gan, Jiaxing Zhang, Baoyuan Wang, et al. 2023. Orca: A few-shot benchmark for chinese conversational machine reading comprehension. *arXiv preprint arXiv:2302.13619*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Somil Gupta, Bhanu Pratap Singh Rawat, and Hong Yu. 2020. Conversational machine comprehension: a literature review. *arXiv preprint arXiv:2006.00671*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Minghui Qiu, Xinjing Huang, Cen Chen, Feng Ji, Chen Qu, Wei Wei, Jun Huang, and Yin Zhang. 2021. Reinforced history backtracking for conversational question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13718–13726.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

S. Reddy, D. Chen, and C. D. Manning. 2019a. CoQA: A Conversational Question Answering Challenge. https://stanfordnlp.github.io/coqa/. Accessed on April 17, 2024.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019b. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Y Sun, S Liu, C Chen, Z Dan, and X Zhao. 2021. Construction of high-quality tibetan dataset for machine reading comprehension. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 208–218.

Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A Fox. 2020. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*.

Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. 2022. Conversational question answering: A survey. *Knowledge and Information Systems*, 64(12):3151–3195.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wenhao Zhuang, Yuan Sun, and Xiaobing Zhao. 2024. Tilamb: 基于增量预训练的藏文大语言模型(tilamb: A tibetan large language model based on incremental pre-training). In *Proceedings of the 23th Chinese National Conference on Computational Linguistics*.

夏天赐and 孙媛. 2018. 基于联合模型的藏文实体关系抽取方法研究. 中文信息学报, 32(12):76–83.

安波and 龙从军. 2022. 基于预训练语言模型的藏文文本分类. 中文信息学报, 36(12):85–93.

李亚超, 江静, 加羊吉, and 于洪志. 2015. Tip-las: 一个开源的藏文分词词性标注系统. 中文信息学报, 29(6):203–207.

龙从军, 刘汇丹, 诺明花, and 吴健. 2015. 基于藏语字性标注的词性预测研究. 中文信息学报, 29(5):211–216.