TiComR：基于提示的藏文对话型阅读理解模型

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

现有的对话型阅读模型在中英文对话型阅读理解任务中表现出色，但由于藏文在语法结构、表达方式等方面同中英文有显著差异，导致这些模型在对藏文对话型阅读理解的对话历史进行建模时存在困难。鉴于此，本文利用当前大模型的优越能力，提出了一种基于提示的对话历史建模方法–TicomR，以解决藏文对话型阅读理解任务中模型性能受限的问题。该方法通过引入基于提示的学习机制，直接在段落文本中添加提示来突显对话历史，而非修改段落标记嵌入，从而在微调过程中实现对对话历史的精确建模，以增强模型对问题的理解能力。实验结果表明，TiComR模型在藏文对话型阅读理解任务上取得了显著的性能提升，并在英文数据集CoQA上也有较好的表现。本文将TicomR开放供研究使用，https://github.com/Tshor/TicomR。

关键词： 藏文；对话型；文本理解；提示学习；大模型

TiComR: A Prompt-based Tibetan Conversational Reading Comprehension Model

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

Existing conversational reading models excel in Chinese and English conversational reading comprehension tasks. However, these models struggle with Tibetan conversational reading comprehension due to significant differences in grammatical structure and expression between Tibetan and Chinese/English. To address this issue, this paper introduces TicomR, a prompt-based conversational history modeling method that leverages the advanced capabilities of current large models. Instead of modifying paragraph token embeddings, TicomR enhances the modeling of conversational history by directly integrating prompts within the paragraph text, allowing for precise fine-tuning and improved understanding of questions.Experimental results demonstrate that the TiComR model significantly improves performance in Tibetan conversational reading comprehension tasks and also performs well on the English CoQA dataset. TicomR is available for research use at https://github.com/Tshor/TicomR.

Keywords: Tibetan , Conversational , Text comprehension , Prompt learning , Large model

第二十三届中国计算语言学大会论文集，第242页–第253页，太原，中国，2024年7月25日至28日。
卷1：主会论文

1 引言

近年来，由于Siri、Alexa和Google Assistant等语音助手平台的广泛应用以及深度学习的进展，对话式问答（Conversational Question Answering, ConvQA）引起了广泛关注 (Rashid et al., 2024)。给定一个文本段落和一段对话，ConvQA的目标是从段落中提取出对话的最后一个问题的答案。目前，大规模的ConvQA基准已经提出，如CoQA (Reddy et al., 2019)、QuAC (Choi et al., 2018)。从技术上讲，ConvQA作为问答（Question Answering, QA）任务的扩展，与传统的单轮问答任务相比，其显著特点是它包含了对话历史，这对于有效建模历史信息提出了挑战。过往研究表明，简单地将对话轮次拼接到输入序列中的方法存在局限 (Gupta et al., 2020)。因此，学者们提出了多种显式建模对话历史的架构组件方案，如HAE (Qu et al., 2019a)、HisBERT (Liu et al., 2020)等。多数研究仅基于特定的CQA基准数据集，如CoQA或QuAC汇报了核心指标。尽管研究人员已在各项基准数据集上取得了显著的成功，但仍然存在许多在以前的ConvQA努力中很少提到的重要障碍：现有的ConvQA数据集主要建立中英等主流高资源语言上，不支持藏文等低资源语言在对话式问答领域的探索；以往的ConvQA模型仅建立在英语语料库上，限制了其他语言ConvQA的发展，如Xlnet (Yang et al., 2019)、SDNet (Zhu et al., 2018)等；藏语独特的语法结构、习惯表达和文化背景，使得主流模型难以迁移，甚至对于LLMs来说也很难完全捕捉到，这通常会导致误解或生成听起来不自然的文本。

以ChatGPT (Ouyang et al., 2022)、Llama (Touvron et al., 2023)等为代表的大语言模型(Large Language Models, LLMs)展现出的理解和生成类人文本能力，显著推动了自然语言处理（Natural Language Processing, NLP）领域的进步与发展 (Moradbeiki and Ghadiri, 2024)。通过最新的研究发现，各种任务之间可能存在紧密的联系，这揭示了一种统一的范式，可以潜在地应用于解决各种QA任务，以模拟它们的共性 (Zhong et al., 2022)。然而，由于大规模藏语训练数据的有限可用性和语言的固有特征，将LLMs直接应用于藏语ConvQA任务面临障碍。

为了应对这些挑战，本文跟随Gekhman等人 (Gekhman et al., 2023)的思想提出了一种依靠大模型能力的基于提示学习的藏文对话式阅读理解模型–TicomR（Tibetan Conversational Comprehension Reader），利用大语言模型的知识和能力来作为基础，微调时在基础文档中添加文本提示，以便突出显示对话历史中以前的答案。实验结果表明，TicomR在TiconvQA (DawaCairen et al., 2024)上表现优异，F1值达到82.19，相比最好的基准有15.69个F1值的相对提升，并在CoQA上也有不错的表现。

本文的主要贡献如下:

(1)本文提出一种利用大模型能力作为基础，并采用提示学习方法进行构建的藏文对话式阅读理解模型–TicomR。通过结合大模型的优势和提示学习的灵活性，TicomR模型能够更有效地捕捉藏文对话的复杂性和特殊性，从而提升对话式阅读理解任务的准确性。

(2)TicomR可以利用少量的数据资源，结合高效低成本的提示学习方法显著提升模型性能，这为藏语等低资源语言的对话型问答任务提供了一个有效的解决方案。

(3)实验结果表明，TicomR在TiconvQA以及CoQA数据集上均表现优异。在TiconvQA上模型最好实验结果获得了82.19%的F1值，相比基线模型提升了15.69%。在CoQA上，得到78.12%的F1，相比基线模型提升了1.52个F1值。此外，本文已将TicomR开源至：https://github.com/Tshor/TicomR。

2 相关工作

机器阅读理解（Machine Reading Comprehension, MRC）任务是NLP领域的一个重要分支，它旨在通过理解文本内容来自动回答关于该文本的问题。MRC任务的核心在于系统能否准确地从给定的文本中检索或生成答案，以响应特定的问题 (Rashid et al., 2024)。这一任务最初由文本检索会议（Text REtrieval Conference, TREC）提出 (Voorhees et al., 1999)，目的是通过查询检索相应的文本段落来解答问题。随着研究的深入和技术的进步，MRC任务已经从最初的简单检索扩展到了更加复杂的对话式问答。与传统的单轮MRC不同，ConvQA涉及多轮对话，系统不仅需要理解当前的问题和上下文，还需要记住之前的对话历史，以便更准确地生成或检索答案 (Rashid et al., 2024)。这种多轮对话的模式更接近人类的交流方式，因此对系统的

理解和记忆能力提出了更高的要求。为了应对ConvQA的挑战，研究者开发了如CoQA (Reddy et al., 2019)，DoQA (Campos et al., 2020)等数据集，这些数据集要求模型理解和利用对话历史来更准确地回答问题。

对话历史建模是对话式问答中的一个主要挑战。早期的解决方案主要依赖于递归神经网络（RNN）和注意力机制的变体来处理序列数据和捕捉长距离依赖关系 (Reddy et al., 2019; Choi et al., 2018; Zhu et al., 2018);另一个趋势是使用基于流的方法，使用对话历史的向量表示在段落中产生潜在的表示 (Huang et al., 2018; Yeh and Chen, 2019; Chen et al., 2021)。然而，随着深度学习的发展，基于Transformer模型的预训练语言模型，如BERT (Devlin et al., 2019)和GPT (Radford et al.,)，已经成为了处理这类任务的主流方法。这些模型通过在大规模语料库上进行预训练，学习到了丰富的语言表示，能够更好地理解和生成自然语言。为了更好地对对话历史进行建模，研究者们尝试了多种方法。例如，FlowQA (Huang et al., 2018)和GraphFlow (Chen et al., 2021)使用每个单词作为图形中的节点，并使用注意机制来表示历史；HAE (Qu et al., 2019a)认为历史基础答案是上下文，这对于现实生活中的对话代理来说是不切实际的；Pos-HAE (Qu et al., 2019b)将历史转向位置视为附加编码；还有基于回溯 (Qiu et al., 2021)和基于查询重写 (Vakulenko et al., 2021)等方法的模型。这些方法通常涉及到对历史信息的编码和更新，以便在生成答案时能够考虑到整个对话的上下文。这种全局视角对于理解和生成连贯、一致的多轮对话至关重要。

在过去的几年中，大型语言模型在NLP的多个领域取得了显著的成就。这些模型依赖于深层的神经网络架构和大规模的预训练数据集，从而在多种NLP任务中展现出卓越的性能。LLMs之所以能够取得这些成果，很大程度上归功于它们在预训练阶段学习到的强大语言表示能力 (Zhao et al., 2023)。LLMs在各种NLP任务中的应用非常广泛，包括但不限于文本生成、情感分析、实体识别和机器翻译等。这些任务通常首先基于预训练的模型进行构建，随后通过微调过程进一步优化模型参数，以更好地适应特定的任务需求。微调过程中，模型在特定任务的数据集上进行训练，从而调整预训练参数以提高任务性能。近期的研究趋势表明，面向提示的微调方法正逐渐受到研究者的关注 (Liu et al., 2023)。这种方法通过在输入文本中插入硬提示标记，如手动设计的标记 (Schick and Schütze, 2021)或自动搜索得到的标记 (Jiang et al., 2020)，来引导模型关注输入文本的关键部分。这些硬提示标记类似于预训练任务中的提示，但它们被重新设计以适应下游任务的目标。然而，考虑到硬提示标记在连续的嵌入空间中可能不是最优解，最新的研究开始探索可调软提示的概念 (Zhong et al., 2022)。软提示通过引入可学习的参数来调整输入文本的表示，从而在模型尺寸较大时实现更加令人满意的性能 (Lester et al., 2021)。与大型预训练语言模型（Pretrain Language Models, PLMs）中的复杂参数相比，软提示提供了一种轻量级且可插拔的解决方案，这为适应新的问答任务提供了更大的灵活性。

与英语等被广泛研究的语言相比，对藏语对话型阅读理解的研究仍处于早期阶段。现有的研究主要集中在单轮抽取式机器阅读理解任务 (Sun et al., 2021)，受限于低资源数据集的稀少、藏语语法规则、文化背景差异等原因，主流模型无法直接迁移至藏文。当前模型在处理高度语境化的问题时，面临着自然语言固有的模糊性挑战，尤其是需要深入挖掘隐含信息，超越表层问题以更深入理解上下文 (Moradbeiki and Ghadiri, 2024)。这些局限性凸显了对研究更加适应或更加稳健的藏语ConvQA模型的迫切需求。

3 TicomR模型

受到当前大语言模型在各项任务上展现出来的优异性能，以及提示学习方法的高效性启发，本文提出了TicomR。这是一种基于提示的对话历史建模方法，该方法通过在段落中的相应位置插入文本提示来突出显示先前对话回合中答案的证据文本，而不是嵌入对话历史的向量表示来对其编码。我们期望能够引导大模型更加聚焦于有价值的对话内容，同时降低现有基于嵌入方法所带来的学习复杂性，进而提升模型在对话处理任务中的性能与效率。

3.1 任务定义

给定一个文本段落P，当前问题Q_k 以及以一系列先前问题和答案的形式存在的对话历史$H_k = (Q_1, A_1, Q_2, A_2, \ldots, Q_{k-1}, A_{k-1})$，一个对话式问答（ConvQA）模型基于文本段落$P$ 和对话历史H_k作为知识源来预测答案A_k。这些答案可以是段落P 内的文本片段（抽取式）或自由格式的文本（生成式）。

3.2 面向藏文对话型阅读理解的提示方法

TicomR采纳了标准的多轮问答模型框架和输入机制，其输入涵盖了当前问题Q_k和段落P以及对话历史H_k。针对每个CQA示例(P, H_k, Q_k)，TicomR会基于先前的对话内容，在段落P中插入文本提示。在抽取式问答的情境下，答案A_k往往是段落P中的某个片段。当给定输入(P, H_k, Q_k)时，TicomR会将段落P转化为带有证据文本高亮显示的版本P^*（通过创建提示并嵌入到P中）。这些提示由定位P中所有历史答案的起始和终止位置组成，并在相应位置插入独特的文本标识符（详见图1中的实例）。随后，经过处理的输入P^*将传递给大模型进行微调，而非原始段落。在生成式问答的场景中，TicomR首先从段落P中提取证据范围，随后生成自由文本形式的答案。因此，最终答案可能并不直接出现在P中。为支持这一情境，TicomR会突出显示历史证据范围（存在于P中的部分），而非直接生成的答案。

为编码对话中的位置信息，答案的标记采用了其逆序转置索引号，即$k-1-j$。这种编码方式反映了历史答案相对于当前问题Q_k的位置，使得模型能够依据它们的顺序来区分不同的历史答案(例如图1中的<1>标引之间的文本即为相对于当前问题Q9的第一轮对话历史，也就是第8轮对话的证据文本)。在CQA对话中，有时会遇到无法回答的问题。对于包含这类问题的文本，TicomR在插入提示之前，会先在P中附加一个“NO ANSWER”的字符串。随后，与常规历史答案相似，每个历史“NO ANSWER”也会被突出显示。例如，请参见图1中的A_9。

总体而言，TicomR相较于以往的提示方法具备以下特征：（1）标记化形式，使用尖括号来标记提示或指示，这种标记化形式使得对话历史更加清晰可辨（例如，“<1>”）。尽管大多数文本提示方法利用由自然语言组成的提示，但我们的提示包含非言语符号，这些符号已被证明对NLP任务的监督具有用处。例如，Aghajanyan等人 (Aghajanyan et al., 2021)通过向输入文本中添加HTML符号展示了结构化预训练的实用性。（2）数字编码，每个标记中的数字同对话历史的轮数关联，引导模型找出当前问题对应的证据文本，以使模型具备当检索定位到多个相同的证据文本时，明确当前问题的证据文本位置。（3）动态性，与大多数预定义提示在输入中位置的提示方法相比 (Liu et al., 2023)，我们的提示会针对每个示例插入到不同的位置。（4）缺失处理，引入<NO ANSWER>标记用于处理证据文本缺失的情况，使得整体提示更具可读性和完整性。这是由于TiconvQA中包含抽取式、是否类以及不可回答类问题，这使得答案和证据文本有可能并不是直接存在于文本中。

3.3 提示设计

如上所述，TicomR在段落中每个历史答案的开始和结束位置插入提示（图1）。提示设计有预定义的标记符号，并包括答案的回合索引（例如，“<1>”）。该设计主要考虑到以下两点：首先，文本提示可以表示对话历史信息；其次，提示在P中的定位有助于引导模型找出当前问题对应的证据文本，以使模型具备当检索定位到多个相同的证据文本时，明确当前问题的证据文本位置。

我们将输入修改为段落和当前问题，而不保留对话历史。这使得模型只能依赖在文本中证据标引来获取历史信息，实验结果表明，TicomR确实编码了来自对话历史的信息，因为显性对话历史的输入与否对于模型性能有显著变化。另外，我们将证据文本在文本中的提示调整为“<>”，即没有转向索引，我们发现，这使得模型性能下降了3.76%，这表明在段落中的提示位置至关重要，并且TicomR的大部分性能增益来自于其提示相对于历史答案的位置。此外，我们将提示策略修改为插入相同数量的符号，但在段落中的相同证据文本但不同位置或随机位置。我们发现模型性能下降了5.08%，表明只有当提示被插入到有意义的位置时，模型才能学会利用这些位置以得到有效的历史表示。

最后，由于文本提示允许轻松注入额外信息，Gekhman等人 (Gekhman et al., 2023)在这个方向上进行了几次初始尝试，将不同类型的信息注入到文本提示中，它们的实验结果表明，提示应该保持简洁符号并在段落中占用小空间的证据。本文尝试使用藏文数字来进行标引，查看同文本一致的提示类型是否有助于提升模型性能，实验结果表明引起的变化并不显著。

3.4 基于大模型的对话型阅读理解微调

大模型微调方法是一种针对预训练模型的技术，主要用于提升预训练模型在特定任务或领域上的性能。本文选择TiLamb (Zhuang et al., 2024)作为基座模型进行藏文对话型阅读理解任务微调及一系列工作。这是一个基于LLaMA2-7B增量预训练的藏文大语言模

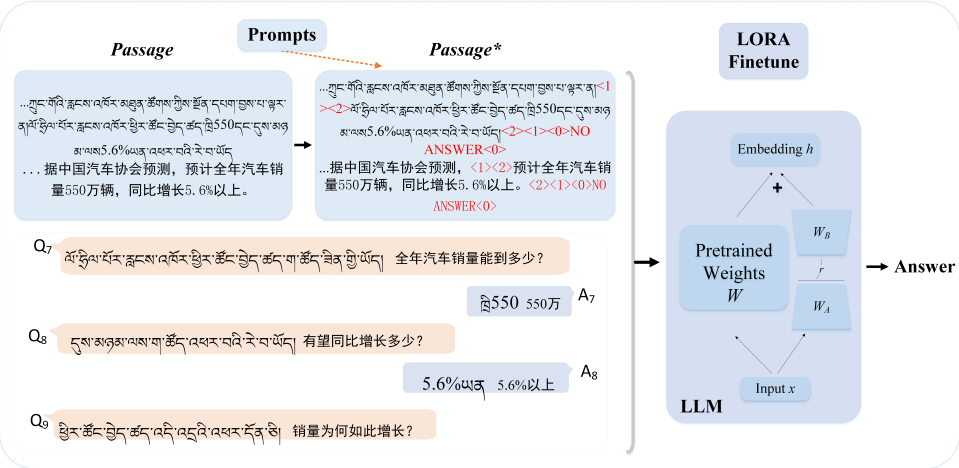

图 1: TicomR中提示突出显示方法

型，它使用了26.43GB的藏文语料进行了增量预训练，并在LLaMA2的基础上扩展了词表，从原有的32000扩充藏文词汇至61221。具体而言，扩展词汇表的过程包括以下步骤：首先，我们使用SentencePiece工具的BPE算法在藏文语料库上训练了一个词表大小为32,000的藏文分词模型；其次，将藏文分词模型与LLaMA原生的分词模型进行合并，通过将藏文tokenizer的tokens添加到LLaMA tokenizer中，创建一个新的tokenizer，该tokenizer能够同时处理英文和藏文文本；最后，排除重复的token后，对词表新增的token进行初始化，并采用均值扩充的方法赋以初值。此外，TiLamb还对LLaMA2-7B原始模型的embedding和lm_head进行了均值扩充初始化。实验结果表明，TiLamb在多个藏文NLP下游任务上展示了出色的性能，例如藏文新闻分类、藏文实体关系分类、藏文机器阅读理解等。TiLamb不仅继承了Llama2的强大能力，而且针对藏文进行了特定的优化和增量预训练，这使得我们的方法能够更有效地应用于藏文对话型阅读理解任务。这一选择是出于两个方面的考虑：首先，尽管LLaMA、Qwen等大模型在中英文等主流语言上表现出色，但并不支持藏文等低资源语言，这限制了我们在藏文对话型阅读理解任务上直接利用主流LLM的能力。其次，TiLamb作为专门为藏文设计的增量预训练大模型，能够更好地捕捉藏文的语法结构和表达方式，从而更准确地建模藏文对话历史。

本文选择LoRA（Low-Rank Adaptation）方法 (Hu et al., 2021)进行TiLamb的藏文对话型阅读理解任务微调。LoRA作为一种高效的大模型微调方法，其核心思想是通过在其权重矩阵中引入低秩矩阵来适应预训练的语言模型。该低秩自适应层初始化随机值，并在微调过程中更新。具体来说，LoRA首先从一个经过预训练的语言模型开始，在其权重矩阵中添加一个低秩适应层。这一低秩适应层以随机初始化的低秩矩阵形式呈现，为模型提供了灵活性和适应性。随后，在新任务或领域的训练过程中仅专注于更新这一低秩适应层，而预训练模型的原始权重则保持不变。这允许模型在不改变其一般知识的情况下有效地学习特定于任务的信息。最终，然后利用经过微调的模型，使用经过调整的模型对来自目标任务或领域的新数据进行预测。通过将适应过程集中在低秩矩阵上，LoRA显著提升了微调的效率，相较于全面微调，大幅减少了计算和内存的开销。在数学上，假设输入是x，输出是h，预训练模型的权重是W_0，降维矩阵是A，升维矩阵是B。采用LoRA微调时，公式如（1）所示：

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{1}$$

其中，ΔW是需要更新的参数，它的分解意味着我们需要用两个较小的LoRA矩阵A和B来表示较大的矩阵。

4 实验评估

4.1 数据集

本文在TiconvQA (DawaCairen et al., 2024)和CoQA (Reddy et al., 2019)上进行了实验，将数据集按照8：2的比例随机划分为训练集、测试集，表1是这些数据集的统计信息。

dataset	passage	Q/A pairs	passage length	turns per passage
TiconvQA	2120	20358	198	9.6
CoQA	8399	127k	271	15.2

表 1: 实验数据集信息

TiconvQA：该数据集是一个藏文对话型阅读理解数据集，通过人工和半自动方法进行构建。TiconvQA拥有来自2120篇地理、人物和新闻三个领域的文本产生的20138轮对话。其中地理和人物领域的文本来自云藏百科，新闻文本来自人民网藏版等权威新闻网站。

CoQA:CoQA数据集收集了来自七个不同领域的8,000个对话，共计126,000个问题，主要来源于CNN等新闻文章。对话构建过程中，采用了两名注释者，一个提问一个回答，都涉及整个上下文。问题是自由陈述，但需要足够的证据和可用的推理。答案是自由形式的，并在文章中强调了相应的基本原理。

4.2 评估指标

为了评估模型的效果，本文使用EM值和F1值两个指标进行评价。

EM值：EM值（Exact Match）是一个关键指标，用于衡量模型预测的答案与标准答案之间的完全匹配程度。它反映了模型在精确捕获问题相关细节和语义方面的能力，是评估模型性能的重要标准之一。高EM值通常意味着模型在理解和回答问题方面具有更高的准确性，具体计算方法如公式（2）。

$$EM = \frac{\text{完全匹配的样本数}}{\text{总样本数}} \tag{2}$$

其中，完全匹配的样本数是指模型输出与参考输出完全一致的样本数量。总样本数是指评估过程中使用的样本总数。

F1值：在对话式阅读理解中，F1值作为评估指标，通过计算精确率和召回率的调和平均数，综合衡量模型性能。它既能体现模型预测的准确性，又能反映模型对正例的覆盖能力，是评估对话式阅读理解模型效果的关键指标，如公式（3）-（5）所示。

Precision（精确度）衡量预测为正例的样本中真正为正例的比例，计算公式为：

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

Recall（召回率）衡量所有真正正例中被模型预测出来的比例，计算公式为：

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

*F1*则是Precision和Recall的调和平均值，用于综合评估模型的性能，计算公式为：

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

其中，真正例（True Positive, TP）表示实际为正例且预测也为正例的样本数，假正例（False Positive, FP）表示实际为反例但预测为正例的样本数，假反例（False Negative, FN）表示实际为正例但预测为反例的样本数。

Parameters	Values
cutoff_len	2048
learning_rate	2e-4
num_tran_epochs	2.0
max_samples	100000
per_devices_train_batch_size	4
gradient_accumulation_steps	4
max_grad_norm	1.0
lora_rank	8
lora_dropout	0.05
resume_lora_training	True

表 2: TicomR模型实验参数具体细节

4.3 实验参数设置

我们对选取的基线模型的实验参数均遵循原始论文的最好设置，在此不做详细描述。对于TicomR的具体实验参数设置如表2所示，所有实验均在2张Tesla V100-PCIE-32G上完成。在表3中我们比较了基线模型在TiconvQA和CoQA上的结果。

4.4 在经典数据集上的表现

本文选择DrQA (Chen et al., 2017)、SDNet (Zhu et al., 2018)、TiBERT (Liu et al., 2022)、TBERT (Dslab-NLP, 2023)等作为基线模型，以上前两者是英文领域经典模型，TiBERT、TBERT是现有常用的藏文预训练语言模型，它们在英文、藏文领域文本分类、情感分析等下游任务上有着出色的表现。具体实验结果表3所示。

（1）DrQA：DrQA是一种基于深度学习的问答模型，通过双向长短期记忆网络（BiLSTM）编码文档和问题，并结合注意力机制。其特点包括端到端的训练、支持多种问答任务、良好的扩展性和高效性，以及通过交互式学习进一步提高性能。

（2）SDNet：SDNet是一种用于端到端问答的神经模型，它通过联合学习问题理解、证据检索和答案生成的过程来提高性能。模型采用了一种层次化的注意力机制，首先对问题进行编码，然后从支持文档中检索相关信息，最后利用这些信息生成准确的答案。SDNet在多个问答数据集上进行了评估，显示出优于传统方法和现有神经模型的性能。

（3）TiBERT：TiBERT是针对藏语自然语言处理任务设计的预训练语言模型。它在大规模藏语训练数据上进行训练，利用Sentencepiece构建了能够覆盖大部分藏语单词的词汇库。TiBERT在文本分类和问题生成等下游任务上展现了出色的性能，并相比经典模型和多语言预训练模型具有优势。该模型的推出为藏语自然语言处理领域的发展提供了有力支持。

（4）TBERT：TBERT（Tibetan-BERT）模型是一种针对藏文自然语言处理任务设计的预训练语言模型。由青海师范大学省部共建藏语智能信息处理及应用国家重点实验室的多拉教授团队与兰州大学开源软件与实时系统教育部工程研究中心共同开发。TBERT旨在解决藏语在自然语言处理领域的数据资源限制和技术挑战，推动藏文信息处理技术的发展。模型基于BERT架构，使用SentencePiece分词器，适用于各种藏语NLP任务。

TicomR在CoQA和TiconvQA两个数据集上均展现出良好的性能。在CoQA数据集上，TicomR的EM和F1值分别达到了50.6%和78.1%。与SDNet相比，TicomR在在F1值上提升了1.5%；与DrQA相比，TicomR在F1值上提升了22.5%。而在TiconvQA数据集上，TicomR的表现亦优于其他模型，其EM值达到了58.9%，F1值达到了82.1%。具体而言，与SDNet相比，TicomR在F1值上提升了35.9%；与DrQA相比，TicomR在F1值上提升了15.6%。

4.5 不同提示方法对模型性能的影响

为了深入探讨不同提示方法对模型性能的影响，我们在实验中尝试了是否添加对话历史(H_k)、对话历史位置空缺提示(<>)、随机位置标引提示(random<>)、藏文数字标引提示(<Tibetan number>)以及对话历史位置相对当前问题逆向标引提示(TicomR)等多种策略，

Model	TiconvQA		CoQA	
	F1 (%)	EM (%)	F1 (%)	EM (%)
人类表现	89.5	80.2	88.8	-
DrQA	66.5	44.5	55.6	46.2
SDNet	46.2	-	76.6	-
TiBERT	40.6	32.7	-	-
TBERT	37.4	21.8	-	-
TicomR	**82.1**	**58.9**	78.1	50.6

表 3: TicomR在不同数据集上的表现

并对比了它们在TiComR模型上的表现。实验结果表明，逆向标引提示方法在TiComR模型上取得了最佳性能。如表4所示，相较于其他提示方法，逆向标引提示在F1和EM指标上均表现出了明显的优势。这种提示方法充分利用了对话的连贯性和上下文信息，有效提升了模型对文本的理解和响应能力。

Prompts	F1	EM
H_k	80.39	56.82
<>	78.43	52.26
random<>	77.11	54.48
<Tibetan number>	81.12	57.40
TicomR	**82.19**	**58.92**

表 4: 不同提示对于模型性能影响

具体而言，尽管对话历史已通过提示方法标引在文章里，但仍然提供原始对话历史有助于模型更好地理解对话轮之间的潜在联系。而只保留<>符号不提供相对位置使得模型无法清晰地定位对应对话历史，虽然简单直接，但可能无法充分捕捉对话的语义信息，导致模型性能有显著下降，这也证明了对话历史的相对位置对于模型理解对话轮之间的关联非常重要。随机文本提示则由于其随机性，无法为模型提供稳定且有效的上下文信息，并有可能导致模型接收到无意义的噪声信息，进一步影响性能。随后，我们尝试了使用藏文数字标引提示来指示对话历史的位置。这种方法相较于前两种提示方法有所改进，但仍然不是最优解。尽管藏文数字为模型提供了位置信息，但可能由于模型对于藏文数字标引并不敏感，只是将其理解为文本信息。相比之下，逆向标引提示能够根据对话的实际情况动态生成提示信息，从而更好地引导模型理解和生成符合语境的响应。这种方法不仅提高了模型的准确率，还增强了模型的鲁棒性，使其能够更好地应对各种复杂的对话场景。

综上所述，不同提示方法对模型性能具有显著影响。在选择提示方法时，我们需要综合考虑对话的连贯性、语义结构以及历史信息的利用方式。通过设计合理的上下文相关提示，我们可以有效提升对话型阅读理解模型的性能。

4.6 消融实验

在本节中，我们进行了一系列消融实验，以验证TicomR模型中各个组件的有效性。实验结果表明，在引入提示后，TicomR模型的性能得到了显著提升，这证明了提示机制在增强模型性能方面的重要作用。具体而言，如表5所示，引入提示后，模型在EM和F1指标上均获得了显著提升。与仅使用微调（Finetune）的模型相比，添加提示（Prompts）的模型在EM和F1指标上分别提高了3.8和5.05个百分点。这表明提示能够有效地帮助模型捕捉和理解对话中的关键信息，从而提高了回答的准确性。

进一步地，当我们综合考虑TicomR模型的所有组件时，其性能得到了进一步的提升。与仅添加提示的模型相比，完整的TicomR模型在EM和F1指标上分别增加了5.68和6.26个百分点。这一结果表明，TicomR模型中的各个组件相互协作，它们共同增强了模型对对话语境的捕捉和

Model	EM	Increase	F1	Increase
+Finetune	53.24	-	75.91	-
+Prompts	57.04	+3.8	80.96	+5.05
TicomR	58.92	+5.68	82.19	+6.26

表 5: TicomR的消融实验

理解能力。

通过消融实验，我们深入了解了TicomR模型中各个组件的贡献。这些结果不仅验证了提示机制的有效性，还为我们提供了关于如何进一步优化模型的宝贵启示。在未来的工作中，我们将继续探索更多有效的提示策略和方法，以进一步提升TicomR模型的性能。此外，值得注意的是，尽管提示机制在增强模型性能方面发挥了重要作用，但可能并不是唯一的因素。在实际应用中，我们还需要综合考虑如模型的架构、训练数据的质量和数量等，以全面提升模型的性能。因此，在未来的研究中，我们将继续探索这些因素对模型性能的影响，并寻求更优化的解决方案。

4.7 不同提示策略下历史长度的影响

在本节中，我们探讨了不同长度的对话历史信息对模型性能的影响。我们将不同长度的对话历史信息与当前问题进行拼接，分别考虑了将当前问题前h轮的对话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=all四种情况。这里的h=all表示将当前问题之前的所有对话历史信息都进行拼接。我们使用了两种不同的提示策略：正向提示和逆向提示，并将它们与原始输入进行了对比。实验结果如表6所示。从表中数据可以看出，无论是采用正向提示和逆向提示，模型在引入对话历史信息后性能均有所提升，这验证了对话历史信息在对话问答中的有效性。

Strategy		h0		h2		h4		h-all	
		F1	EM	F1	EM	F1	EM	F1	EM
original		73.32	51.17	75.91	53.24	76.08	53.46	75.87	53.01
prompt	reverse	79.51	56.52	80.96	57.84	**82.19**	**58.92**	81.95	56.55
	forward	79.42	55.89	80.69	57.18	81.46	58.12	80.43	56.21

表 6: 不同提示策略下的实验结果

进一步地，我们观察到对于逆向和正向两种提示策略，它们在利用对话历史信息方面表现出不同的特点。具体而言，逆向提示策略在多数情况下优于正向提示策略。这可能是因为逆向提示更能有效地捕捉对话历史中的关键信息，因为它按照从最近到最远的顺序标引对话历史，使得模型更能够关注到与当前问题最为相关的历史信息。另外，我们观察到当h取值为4时，模型性能达到最佳。这表明适当数量的历史信息对于模型性能至关重要。过多的历史信息（如h=all）可能导致模型在处理时引入噪声，而过少的历史信息则可能无法提供足够的上下文信息。因此，选择合适的h值对于平衡信息量和模型性能至关重要。

5 总结与展望

本文针对藏文对话型阅读理解任务中模型性能不足的问题，提出了一种新型的基于提示的藏文对话型阅读理解模型——TiComR。通过结合当前大模型的优异能力和基于提示的学习机制，TiComR显著提升了在藏文对话型阅读理解任务上的性能。实验结果表明，TiComR模型在TiconvQA和CoQA数据集上都取得了优异的表现，未来可以进一步扩充训练数据集、优化模型结构，并将TiComR模型应用到其他低资源语言的对话型阅读理解任务中，以验证其通用性和适用性。

致谢

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

参考文献

Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2021. Htlm: Hyper-text pre-training and prompting of language models. In *International Conference on Learning Representations*.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. Doqa–accessing domain-specific faqs via conversational qa. *arXiv preprint arXiv:2005.01328*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2021. Graphflow: exploiting conversation flow with graph neural networks for conversational machine comprehension. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1230–1236.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

DawaCairen, PengmaoCairang, and Yuan Sun. 2024. 面向对话式阅读理解的高质量藏语数据集构建(construction of high-quality tibetan language dataset for conversational reading comprehension). In *Proceedings of the 23th Chinese National Conference on Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dslab-NLP. 2023. Tibetan-plm. https://github.com/Dslab-NLP/Tibetan-PLM.

Zorik Gekhman, Nadav Oved, Orgad Keller, Idan Szpektor, and Roi Reichart. 2023. On the robustness of dialogue history representation in conversational question answering: A comprehensive study and a new prompt-based method. *Transactions of the Association for Computational Linguistics*, 11:351–366.

Somil Gupta, Bhanu Pratap Singh Rawat, and Hong Yu. 2020. Conversational machine comprehension: a literature review. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2739–2753.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2018. Flowqa: Grasping flow in history for conversational machine comprehension. In *International Conference on Learning Representations*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Chuang Liu, Deyi Xiong, Yuxiang Jia, Hongying Zan, and Changjian Hu. 2020. Hisbert for conversational reading comprehension. In *2020 International Conference on Asian Language Processing (IALP)*, pages 147–152. IEEE.

Sisi Liu, Junjie Deng, Yuan Sun, and Xiaobing Zhao. 2022. Tibert: Tibetan pre-trained language model. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2956–2961. IEEE.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Pardis Moradbeiki and Nasser Ghadiri. 2024. Perkwe_coqa: enhance persian conversational question answering by combining contextual keyword extraction with large language models. *arXiv preprint arXiv:2404.05406*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

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.

Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019a. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1133–1136.

Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W Bruce Croft, and Mohit Iyyer. 2019b. Attentive history selection for conversational question answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1391–1400.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

Muhammad Shihab Rashid, Jannat Ara Meem, and Vagelis Hristidis. 2024. Normy: Non-uniform history modeling for open retrieval conversational question answering. *arXiv preprint arXiv:2402.04548*.

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

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Yuan Sun, Chaofan Chen, Sisi Liu, and Xiaobing Zhao. 2021. Ti-reader: 基于注意力机制的藏文机器阅读理解端到端网络模型(ti-reader: An end-to-end network model based on attention mechanisms for tibetan machine reading comprehension). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 219–228.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 355–363.

Ellen M Voorhees, Dawn M Tice, et al. 1999. The trec-8 question answering track evaluation. In *TREC*, volume 1999, page 82.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yi-Ting Yeh and Yun-Nung Chen. 2019. Flowdelta: Modeling flow information gain in reasoning for conversational machine comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 86–90.

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*.

Wanjun Zhong, Yifan Gao, Ning Ding, Yujia Qin, Zhiyuan Liu, Ming Zhou, Jiahai Wang, Jian Yin, and Nan Duan. 2022. Proqa: Structural prompt-based pre-training for unified question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4230–4243.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.

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*.