基于数据增强的藏文机器阅读有难度问题的生成

旦正错[1,3] 陈龙[1,3] 邓俊杰[1,3] 庞仙[2,3] 孙媛[1,3,4,*]
[1]中央民族大学 信息工程学院，北京 100081
[2]中央民族大学 中国少数民族语言文学学院
[3]国家语言资源监测与研究少数民族语言中心
[4]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

问题生成是机器阅读理解数据集构建的子任务，指让计算机根据给定有（无）答案的上下文，生成流利通顺的问题集。在中英文领域，以端到端为基础的问题生成模型已经得到了很好的发展，并且构建了大批高质量的问答对。但是在低资源语言（藏文）领域，以机器阅读理解、智能问答系统为代表的数据驱动型任务中仍然普遍存在数据量较少和问答对过于简单的问题。因此，本文提出了三种面向藏文机器阅读的有难度问题的生成方法:（1）基于藏文预训练语言模型进行掩码、替换关键词生成不可回答问题。（2）根据相似段落的问题交叉生成不可回答的问题。（3）根据三元组生成具有知识推理的问题。最后，本文在构建的数据集上进行了实验，结果表明，包含不可回答、知识推理等类型的机器阅读理解数据集对模型的理解能力提出了更高的要求。另外，对构建的不可回答问题，从数据集的可读性、关联性和可回答性三个层面验证了数据集的质量。

关键词: 藏文；不可回答；有难度；数据集；机器阅读理解

Difficult Question Generation of Tibetan Machine Reading Based on Data Enhancement

Zhengcuo Dan[1,3] Long Chen [1,3] Junjie Deng [1,3] Xian Pang [2,3] Yuan Sun [1,3,4,*]
[1] School of information engineering, Minzu University of China, Beijing 100081
[2] School of Chinese Ethnic Minority Languages and Literature, Minzu University of China
[3] National Language Resources Monitoring and Research Center for Minority Languages
[4]Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE
*Corresponding author: Yuan Sun
tracy.yuan.sun@gmail.com

Abstract

Question generation is a sub task of constructing machine reading comprehension datasets, aimed at enabling computers to generate fluent question sets based on the context of given (no) answers. In the field of Chinese and English, generative models based on end-to-end have been well developed, and a large number of high-quality question and answer pairs have been constructed. However, in the field of low resource language (Tibetan), there are problems of less data and too simple Q&A pairs in data-driven tasks such as machine reading comprehension and intelligent question answering. Therefore, this paper proposes three methods to generate difficult questions for Tibetan machine reading. (1) Mask and replace keywords based on a Tibetan pre-trained language model to generate unanswerable questions. (2) Generate unanswerable questions based on question exchange in similar paragraphs. (3) Generate questions with knowledge reasoning based on triples. Finally, this paper conducts experiments on the constructed dataset, and the results show that machine reading comprehension datasets containing unanswerable, knowledge reasoning, and other types

第二十二届中国计算语言学大会论文集，第164页-第173页，哈尔滨，中国，2023年8月3日至5日。

put forward higher requirements for the model's understanding ability. In addition, for the constructed unanswerable questions, the quality of the dataset was verified from three aspects: readability, relevance, and answerability.

Keywords: Tibetan , Unanswerable , Difficult , Data set , Machine reading comprehension

1 引言

机器阅读理解（MRC）指机器根据给定的上下文回答相关问题，测试机器对自然语言的理解程度。早期的机器阅读理解依赖于人工制定的规则或基于统计学习模型，存在可移植性差、人工成本高、产生周期长的问题，很难在实际中广泛应用。近年来，大规模、高质量的数据集极大地推动了机器阅读理解的发展。(Hirschman et al., 1999)等人第一次构建了面向机器阅读理解的数据集，包括3-6年级的阅读材料和简单的5W问题。随后出现了面向多项选择式机器阅读理解的数据集MCTest(Richardson et al., 2013)、RACE(Lai et al., 2017)；面向完形填空式机器阅读理解的数据集Children's Book Test(CBT)(Hill et al., 2016)、CNN&Daily Mail(Hermann et al., 2015)；面向区间答案式机器阅读理解的数据集SQuAD(Rajpurkar et al., 2016)和面向自由问答式机器阅读理解的数据集MARCO(Nguyen et al., 2016)、DuReader(He et al., 2018)。随着这些大规模数据集的创建与应用，如R-Net(Wang et al., 2017)、BiDAF(Seo et al., 2016)等模型相继被提出并在机器阅读理解任务上取得不错的效果，目前在SQuAD数据集上最好的模型成绩达到了95.71[0]，超过人类的表现。

根据上述数据集训练出来的模型容易受到对抗样本的攻击，如在原先的数据集中加入根据文本内容所产生的干扰片段(Jia and Liang, 2017)、删除问题或片段的重要部分以把当前问题变得不可回答(Mudrakarta et al., 2018)之后模型仍然能根据文章给出合理但不正确的答案。针对此类问题，(Rajpurkar et al., 2018)等人提出了包含"不可回答"问题的数据集SQuADRUN，在原来的SQuAD十万个问题——答案对的基础上，新增了超过五万个由人类众包者设计的无法回答的问题。此数据集让模型先判断当前问题是否存在答案，然后确定答案。这类任务也可以让模型更加符合人类做阅读理解的习惯和思维方式，目前，SQuADRUN上表现最好的模型成绩为93.21[1]，超过人类89.45的表现。

在低资源语言（藏文）领域，TibetanQA(Sun et al., 2021)是面向藏文抽取式、可回答的机器阅读理解数据集，该数据集规模可观、涵盖内容比较全面，但是无法满足不可回答问题、基于多个信息进行多步推理、加入外部知识等藏文机器阅读理解任务的需求。生成大规模有难度机器阅读理解数据集的研究还处于初步阶段，主要原因在于：（1）通过众包的形式从海量无结构化文本数据中构建相应数据集存在人工成本过高、构建周期过长、质量难以掌控的问题。（2）藏文属于黏着语，藏文中的虚词往往与实词组合的形式出现，如"ཡི་"（的），导致词与词之间没有明确的划分，因此，目前中英文领域的模型无法直接套用在藏文机器阅读理解任务上。针对以上问题，本文提出了三种面向藏文机器阅读理解的有难度问题的自动生成方法，并在相关的实验上验证了数据集的质量。本文的主要贡献如下：

（1）本文通过三元组的隐式实体关系链，提出基于三元组的知识推理问题生成方法，并将其与基于语法规则生成的简单问题进行了比较。

（2）本文通过众包的形式构建了面向藏文机器阅读理解不可回答问题数据集，包含2,200对问答对，并使用藏文预训练语言模型，提出了基于掩码、替换关键词的藏文机器阅读理解不可回答问题的生成方法。

（3）本文通过计算藏文机器阅读理解数据集TibetanQA的段落相似度，设置上下限阈值实现藏文机器阅读理解不可回答数据集的增广。

2 相关研究

问题生成的方法分为基于规则的问题生成和基于神经网络的问题生成。基于规则的问题生

[0]https://paperswithcode.com/sota/question-answering-on-squad11

[1]https://paperswithcode.com/sota/question-answering-on-squad20

成主要利用句法分析和知识库的辅助制定相关的规则，将陈述句改为疑问句来生成问题。最新研究表明，在非常成熟的第三方语义资源和强大的句法分析技术的支撑下，基于规则的问题生成效果优于神经网络模型(Dhole and Manning, 2020)，但由于不同语言、同语言不同领域之间的差异性，规则移植性差，难以扩展，并且人为制定的规则限制了生成问题的多样性。

为了减少人力和缩减构建周期，(Du et al., 2017)等人和(Zhou et al., 2018)等人首次提出基于神经网络模型的问题生成研究。(Du et al., 2017)等人提出了一种基于全局注意力机制的序列学习模型来生成问题。之后，众多学者从不同角度和侧重点对端到端的问题生成展开研究。为了解决问题生成中疑问词与答案类型不匹配、复制机制提取的片段与答案词不相关的问题，提出了将词法、词汇特征(Zhou et al., 2018)、答案信息(Wang et al., 2020)、答案的位置信息(Sun et al., 2018)等各种特征作为输入来提升模型性能。但是，将已知答案信息作为输入特征会使模型自动生成的问题中可能包含目标答案，因此，(Kim et al., 2019)等人使用[mask]标记文本中的答案词，在分开的答案词信息中捕获关键信息，最后采用检索式词生成器(Ma et al., 2018)生成完整的、不包含答案词的问题。比起句子级问题生成，段落级问题生成(Zhao et al., 2018)由于融入了更多的语义信息，其生成的问题质量往往更好。

自预训练语言模型（例如BERT(Kenton and Toutanova, 2019)）及其变体被提出，在机器阅读理解等自然语言处理的下游任务中，其表现远超序列到序列的神经网络模型，但对于不可回答、多跳、加入外部知识等机器阅读理解的复杂任务，模型还是无法做出强有力的判断。因此，(Zhu et al., 2019)等人把SQuADRUN作为不可回答问题生成模型的训练数据，pair-sequence（Pair2Seq）作为问题生成模型，自动生成面向机器阅读理解不可回答问题集。除此之外，在知识推理数据集方面，出现了给每个文档附带多个相关文档的TriviaQA(Joshi et al., 2017)、利用知识图谱构建的QAngaroo(Welbl et al., 2018)、基于多个文档且问题不局限于任何已有的知识库或知识模式的HotpotQA(Yang et al., 2018) 等多跳数据集。

目前，在中英文领域，已经出现了很多机器阅读理解不可回答、知识推理、加入外部知识等的复杂数据集及相关研究。但是在低资源语言（藏文）中，其相关研究还处于初步阶段，因此，本文提出了三种低资源语言（藏文）的不可回答和知识推理的机器阅读理解数据集增广方法，以促进低资源语言（藏文）机器阅读理解的发展。

3 模型架构

3.1 知识推理有难度问题的生成

基于深度学习的机器阅读理解模型已经取得了很大的进步，但是与人类相比，其理解能力有四个方面的不足，主要表现在推理能力弱、可解释性差、缺少外部知识、答案可塑性差。本文通过提取文本中包含的三元组实体的隐式关系，构建了基于三元组的知识推理数据集，并与基于规则生成的简单数据集进行了比较。

3.1.1 基于规则的简单问答对生成

为了检验知识推理问题集的质量，本文根据三元组、三元组显式关系的同义词以及相关的藏文格助词添接语法生成了基于规则的藏文机器阅读理解简单问答对，其构建方法分为三元组提取及匹配，三元组关系的同义词统计，基于规则的问题生成。

1、三元组提取及匹配

王丽客等人构建了103,509条藏文知识库(王丽客et al., 2021)，本文使用TibetanQA的文本与其对齐，没有对齐到的文本，提取并标注其适合的三元组，得到<实体1，关系，实体2，文章>格式的数据，如<ལུའུ་ཨུན།，མཚན་དངོས་，ཀྲིག་རྒྱུའུ་རེན་，ལུའུ་ཨུན། ཁོང་གི་མཚན་དངོས་ལ་ཀྲིག་རྒྱུའུ་རེན་ཞུ་བ་རེད། གསར་བཟེ་པ་ཡིན།>(鲁迅，原名，周树人，鲁迅，他的原名叫周树人，革命家)。

2、三元组关系统计及规则制定

对于对齐得到的三元组，即实体和关系，为了减少因关系出现次数太少而产生的噪音，筛选出现次数最高的前4个关系作为制定规则的依据，包含国家、出版物、出生日期、死亡日期，共有1,846条数据，最高出现次数为892，最低出现次数为24。另外，统计三元组关系的同义词，4种关系共11个不同的同义词。最后利用实体、实体关系的同义词得到藏文简单问答对，数据示例如表1所示。

表1中，根据实体1、关系和语法规则生成简单的问题集，而实体2作为生成问题的答案。另外，对于同一组三元组，实体关系的同义词不同，其生成的问题也不同。文中的关

实体1	关系	实体2	生成的问句
ལུའུ་ཞུན། (鲁迅)	Mother	ལུའུ་རུའི། (鲁瑞)	ལུའུ་ཞུན་གྱི་ཡུམ་ཚེན་ནི་སུ་ཡིན།, ལུའུ་ཞུན་གྱི་མ་ཡུམ་ནི་སུ་ཡིན། (鲁迅的母亲是谁？)
ལུའུ་ཞུན། (鲁迅)	Birthday	1881.9.25	ལུའུ་ཞུན་ནི་དུས་ནམ་ཞིག་ལ་སྐྱེ་འཕྲུང་བ་ཡིན། (鲁迅是什么时候出生的？)

表 1. 基于规则生成的简单问题

系Mother有“ཡུམ།”,“ཡུམ་ཚེན་”,“ མ་མ་ ”等同义词，可以生成ལུའུ་ཞུན་གྱི་ཡུམ་ཚེན་ནི་སུ་ཡིན།或者ལུའུ་ཞུན་གྱི་མ་ཡུམ་ནི་སུ་ཡིན། 等不同的问题。

3.1.2 基于三元组生成的多跳阅读理解数据集

三元组是表示文本结构最常用的一种方法。本文从原始段落中抽取三元组，使用图的节点表示三元组的实体，连线表示实体间的关系，若为实线则表示该实体间的关系是显式可控，若为虚线，则表示实体间的关系为隐式不可控。根据实体间的可推理路径，构建了基于三元组多跳的知识推理数据集，阅读此类数据集需要根据当前段落提供的线索，进行2-4的跳级检索才能得到问题的答案，其数据示例如图1所示。

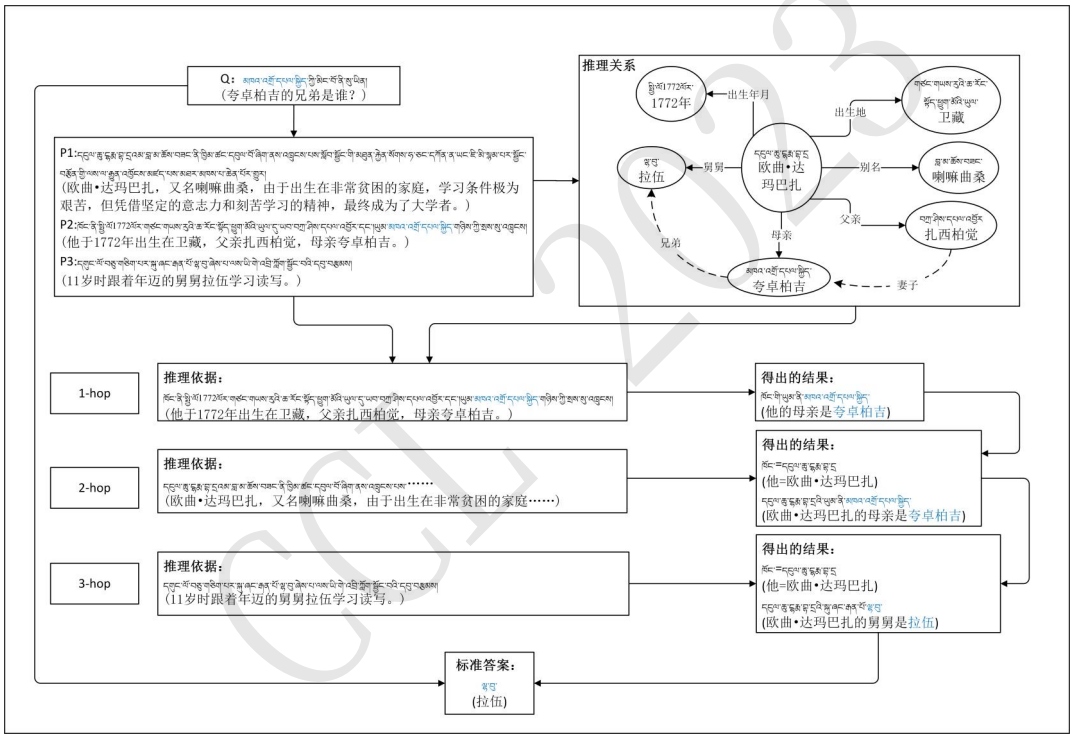

图 1. 基于三元组的知识推理问题生成方法

图1中，问题མཚན་འགོ་དཀར་སྐྱིད་ཀྱི་མིང་ཙོ་ནི་སུ་ཡིན། （夸卓柏吉的兄弟是谁？），在文中没有直接的答案。根据文章，在第一、二段找出实体དཏལ་ཆུ་དཔའ་རྒྱ་དོ། （欧曲·达玛巴扎）与མཚན་འགོ་དཀར་སྐྱིད། （夸卓柏吉）的显式关系为母亲，第三段中找出实体དཏལ་ཆུ་དཔའ་རྒྱ་དོ། （欧曲·达玛巴扎）与ལ་ཕུ（拉伍）的显式关系为舅舅，从而推断出དཏལ་ཆུ་དཔའ་རྒྱ་དོ། （欧曲·达玛巴扎）的舅舅ལ་ཕུ （拉伍）是དཏལ་ཆུ་དཔའ་རྒྱ་དོ། (欧曲·达玛巴扎)的母亲མཚན་འགོ་དཀར་སྐྱིད་(夸卓柏吉)的兄弟。

3.2 不可回答问题的生成

通常，机器在做阅读理解时，需要根据文档判断出当前问题是否可答后才进行下一步的答案提取工作。不可回答数据集是实现这一任务的基础。本文提出了两种藏文机器阅读理解不可回答数据集的构建方法。

3.2.1 基于相似度计算的不可回答问题的生成

一词多义问题是机器阅读理解、机器翻译等自然语言处理下游任务中的难点。藏文中，在不同的语境下，ཉི་མ་(太阳)，མེ་དོག་ (花朵)表示人名，ནམ་མཁའི་ཆོར་ཐུ་ (空中珍宝)、པདྨའི་གཉེན་ (莲花之友)表示太阳。这种情况下传统的编辑距离计算不出这些词语的远近关系，但是这也给机器阅读理解不可回答数据集的构建提供了新的思路。

本文计算了TibetanQA的段落相似度，相似度在0.6-0.9之间的段落在书写层面被认为具有一定的相关性，因此，互相替换该段落的问题集，得到基于相似度计算的藏文机器阅读理解不可回答数据集，数据实例如图2所示。

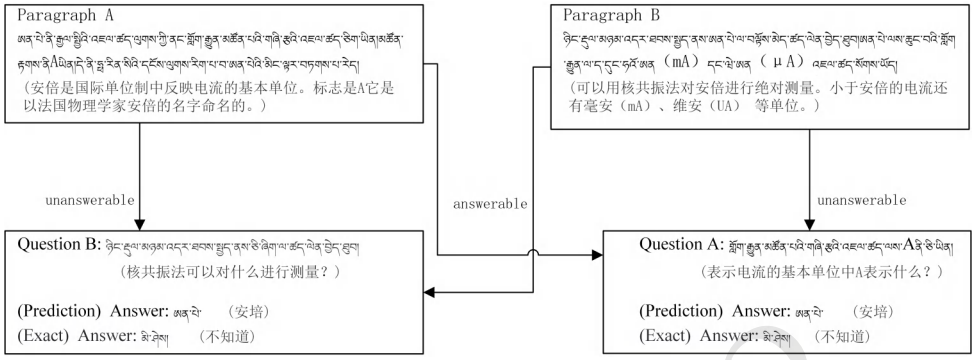

图 2. 基于相似度计算的不可回答问题生成方法

图2中，Question A（B）是Paragraph A（B）的可回答问题，通过计算两个段落的相似度，具有相关性的段落对应的问题集交换并进行人工校对，得到了Paragraph A（B）的不可回答问题Question B（A）。

3.2.2 基于藏文预训练语言模型的不可回答问题的生成

前期，通过众包的形式构建了2,200对面向藏文机器阅读理解不可回答数据集。为了避免下游任务的模型通过简单的启发式搜索或单词匹配的方式在文中找到当前问题的答案，在问题构建过程我们遵循了以下三个原则：（1）对于每篇文章，问题的最佳数量定为1到10；（2）问题集没有相关的答案，但是有看似合理的答案；（3）问题与原上下文内容高度相关。另外，验收时，我们删除了问题数量小于100的创建者提供的问答对，有效避免了没有全面理解此类任务而创建问题的人为噪音。其数据示例如图3所示，其中包含段落、根据该段落提出的不可回答问题和取自该段落的看似合理的答案。

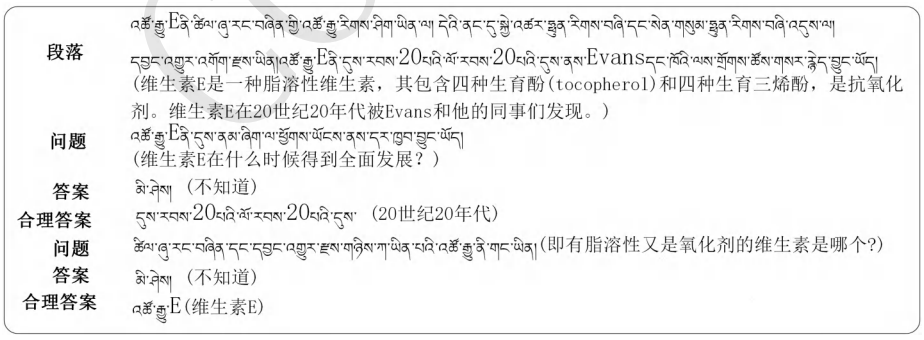

图 3. 通过众包构建的不可回答数据集样例

为了在短周期内获得更具有挑战性的数据集，本文利用藏文预训练模型TiBERT(Liu et al., 2022)对可回答问题进行掩码，替换关键词的方式自动生成不可回答问题，其过程分为确定可回答问题集的关键词、对应文本中的关键词进行掩码和预测，用预测出的关键词替换问题集的关键词生成不可回答问题，数据示例和构建过程如图4所示。

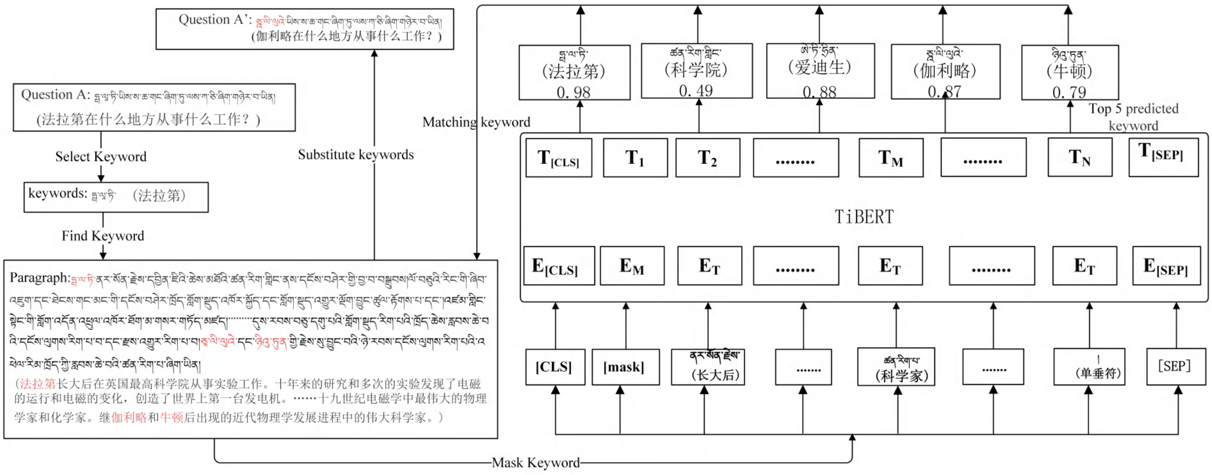

图 4. 基于TiBERT的不可回答问题生成方法

1、确定可回答问题集的关键词

本文使用了TibetanQA中的问题集。注意到问题集中包含许多人名、地名、组织机构名等专有名词，先用sentence piece(Kudo and Richardson, 2018)对语料库进行一体化分词并随机掩码，结果并不理想，主要原因是文本内容涵盖广泛的主题，每个主题的句子结构有比较鲜明的对比。为了正确提取当前问题中的关键词信息，按照文章主题将数据集分为更细粒度的子集，包括人物、时间、科学等类别。本文筛选TibetanQA问题集及相关段落中包含人物名字的数据，提取问题集中的人物名字作为关键词。如图4的Question A:ཧྥ་ལ་ཏི་ཡིས་ས་ཆ་གང་ཞིག་ཏུ་ལས་ཀ་ཅི་ཞིག་གཉེར་བ་ཡིན།（法拉第在什么地方从事什么工作？)中将“ཧྥ་ལ་ཏི་”(法拉第)标为该问题的关键词信息。

2、关键词掩码，预测并生成不可回答问题

确定了问题集中的关键词，接着使用[mask]对文本中的关键词进行掩码，最后使用藏文预训练语言模型TiBERT将其预测，输出排名前五的预测结果。

为了保证得到的问题集是不可回答且具有一定的难度，一方面，TiBERT预测出的关键词不得等同于掩码之前的关键词以及文中表明的该关键词的同义词，另一方面，TiBERT预测出的关键词需跟当前段落有一定的关联，避免模型根据单词重叠等简单的方式就能判断出问题的可回答性。为此，本文将模型预测出的五个关键词返回到问题及相关段落中进行匹配，过滤掉与关键词相同、不存在于当前段落中的预测值。最后，根据藏文格助词语法的添接规则，使用新预测到的关键词替换可回答问题集中的原关键词，产生新的不可回答问题。

如图4所示，[mask]原文中的“ཧྥ་ལ་ཏི་”（法拉第）时按照概率输出的预测结果分别为ཧྥ་ལ་ཏི་（法拉第）、ཨེ་ཏི་སིན་（爱迪生）、ཀླ་ལི་ལུའེ་（伽利略）、ཉིའུ་ཏུན་（牛顿）、ཚན་རིག་གླིང་（科学院）。其中ཧྥ་ལ་ཏི་（法拉第）、ཨེ་ཏི་སིན་（爱迪生）是预测值最高的两个输出，但是ཧྥ་ལ་ཏི་（法拉第）是原问题中的关键词，而ཨེ་ཏི་སིན་（爱迪生）不存在于当前文本中，用前者替换问题中的关键词没有意义，用后者关键词替换的新问题过于简单，模型根据单词重叠就能得出问题的准确答案。去除两者干扰项，最终得到的预测值为ཀླ་ལི་ལུའེ་（伽利略），替换可回答问题Question A中的关键词ཧྥ་ལ་ཏི་（法拉第），产生新的不可回答问题Question A':ཀླ་ལི་ལུའེ་ཡིས་ས་ཆ་གང་ཞིག་ཏུ་ལས་ཀ་ཅི་ཞིག་གཉེར་བ་ཡིན།（伽利略在什么地方从事什么工作？）。

4 实验结果与分析

4.1 实验数据集

本文在TibetanQA和构建的各类数据上进行了实验。将数据按照8:2的比例分为训练集和测试集，分布如表2所示。

TibetanQA(Sun et al., 2021)：采用众包的形式构建的数据集，包含20,000对藏文机器阅读理解可回答问答对和1,513篇文章，文本数据选自云藏百科。

数据集		问答对		
		训练集	测试集	总
推理问题	TibetanQA	16,000	4,000	20,000
	TibetanQA+MultiHop	16,740	4,183	20,923
	TibetanQA+Trip	17,477	4,369	21,846
不可回答问题	TibetanQA+Unanswerable	17,760	4,440	22,200
	TibetanQA+Mask	16,800	4,200	21,000
	TibetanQA+Sim	16,800	4,200	21,000

表 2. 实验所用的数据集

TibetanQA+MultiHop：在TibetanQA中加入了根据三元组多跳的形式构建的知识推理数据集。

TibetanQA+Trip：在TibetanQA加入了三元组及规则生成的简单问答对。

Unanswerable：本文采用众包的形式构建的不可回答数据集，包含2,200对问答对。

TibetanQA+Mask：在TibetanQA加入根据藏文预训练语言模型TiBERT进行掩码，替换关键词的方式生成的藏文机器阅读理解不可回答数据集。

TibetanQA+Sim：在TibetanQA加入了根据可回答数据集相似段落的问题交叉产生的藏文机器阅读理解不可回答数据集。

4.2 实验结果

本文使用(Liu et al., 2022)等人提出的TiBERT在TibetanQA及构建的各类数据集上进行了实验，使用EM和F1值对实验结果进行了评价。

4.2.1 推理问题集的实验结果分析

TibetanQA+Trip、TibetanQA+MultiHop以及TibetanQA数据集在TiBERT上的实验结果如表3所示。

数据集	EM	F1
TibetanQA	53.2	73.4
TibetanQA+MultiHop	47.6	69.9
TibetanQA+Trip	52.9	73.6

表 3. TiBERT在藏文MRC可回答数据集上的实验结果

表3中，TiBERT在TibetanQA数据集上的EM和F1值分别为53.2%和73.4%，根据规则生成的简单问答对上的EM和F1值分别为52.9%，73.6%，其EM值比前者下降了0.3%，而F1值却提高了0.2%，总体对模型的影响较小。其原因如下：该数据集包含的关系较少、生成的问题种类不够丰富、关系词及其同义词具有比较鲜明的特点，如表示母亲的关系词及其同义词“ཨ་མ་”，“ཨ་ཡུམ་”都包含表示女性的词“མ”，这使得模型很容易识别当前问题，从而精准找到文中的答案区间。

采用三元组多跳形式生成的知识推理型数据集在TiBERT上的EM值和F1值分别为47.6%和69.9%，比TibetanQA数据集上的表现下降了5.6%和3.5%。其主要原因是根据多跳三元组产生的知识推理型数据集对模型的理解能力提出了更高的要求，回答此类数据集的问题，模型需要有一定的知识推理能力，无法根据单纯的启发式搜索或者加入以同义词为主的外部知识来获取答案。

4.2.2 不可回答的实验结果分析

本文使用三种方法构建的不可回答数据集TibetanQA+Unanswerable、TibetanQA+Mask、TibetanQA+Sim以及TibetanQA数据集在TiBERT上的实验结果如表4所示。

由表4得知，加入不可回答、知识推理等具有难度的数据集在TiBERT上的结果都呈下降趋势，表明机器阅读理解数据集的不同类型和难度会给模型带来不同程度的影响，有难度的数据集对模型的鲁棒性提出了更高的要求。

数据集	EM	F1
TibetanQA	53.2	73.4
TibetanQA+Unanswerable	50.1	72.6
TibetanQA+Mask	50.1	72.1
TibetanQA+Sim	51.6	73.5

表 4. TiBERT在不可回答数据集上的实验结果

TiBERT在人工构建的不可回答数据集TibetanQA+Unanswerable上的EM值和F1分别为50.1%，72.6%，比TibetanQA上的表现分别下降了3.1%和0.8%，对模型性能产生较大的影响。这也验证了人工方式构建的数据集在答案的不可回答性、答案与文章的相关性和合理答案的选择上具有明显的优势。

根据相似段落的问题交叉产生的不可回答数据集的EM值为51.6%，比TibetanQA数据集下降了1.6%，而其F1值为73.5%，比TibetanQA数据集高出0.1%，对模型的影响较小。表明除了数据量的可控因素之外，不可回答数据集中文章与问题的关联性是影响模型的因素之一。

根据TiBERT进行掩码,替换关键词方法生成的不可回答数据集在TiBERT上的EM值和F1值分别为50.1%，72.1%，比TibetanQA数据集下降了3.1%和1.3%，对模型性能的影响最大，表明将文章根据主题分为更细粒度的子集并进行关键词掩码和替换时产生的效果更好。

另外，本文在每类不可回答数据集中随机选择10%的样本，邀请三组藏族同学根据可读性、关联性和不可回答性三个维度对其进行打分，累计最高为3分，最差为0分。取三种指标的平均值作为最终结果，如表5所示。

可读性：数据集中语法的添接规则、疑问词的使用等书写内容，正确标为1，否则标为0，目的是为了保证数据集书写的规范；关联性：当前问题类型与文本类型是否匹配，即若文本内容是人物介绍类，而问题内容是景物或者其他与人物介绍完全不相关的标为0，否则标为1，其目的是为了避免模型以单词匹配等简单的方式识破不可回答问题；可回答性：当前问题根据给出的文本是否可答，如果不可回答标为1，否则标为0，其目的是检测生成问题的不可回答性。

指标 / 类型	可读性(%)	关联性(%)	不可回答性(%)	平均值(%)
Unanswerable	0.97	0.96	0.85	0.93
Mask	0.75	0.63	0.74	0.71
Sim	0.99	0.44	0.61	0.68

表 5. 不可回答数据集的人工评价结果

表5中，通过众包构建的unanswerable数据集上三种指标的平均值达到了93%，而根据预训练语言模型生成的不可回答数据集Mask和根据相似段落的问题交叉生成的数据集Sim在三种评价指标上的平均值只有71%和68%，比unanswerable数据集分别下降22%和25%。数据表明，机器自动生成的数据集质量还有进一步的发展空间。其主要原因如下：预训练语言模型在生成问题时，将原文中[mask]的人名部分预测成一个代词或者关联性不大的另一个人名，使得生成的问题没有明确的主语。藏文属于黏着语，因此，预测出的新词往往伴随着不同的格助词，这对于根据格助词的添接法则生成问题的规则非常不友好，使得生成的新问题出现语法错误和重复的问题，相似段落交替而生成的问题的可读性指标达到99%，因为相似段落的计算是在TibetanQA的基础上完成，而TibetanQA问题集的语法和疑问词的使用等书写较为规范。但是该类数据集的问题和段落相关性不大，导致其生成的问题与文章内容不相关，机器根据单词重叠的启发式搜索变能分辨当前问题不可答。

5 总结

本文提出了三种面向藏文机器阅读理解的数据增广方法，并且构建了对应的数据集。为了检验数据集的质量，利用藏文预训练语言模型在构建的不同类型的数据集上进行实验，并对藏文机器阅读理解不可回答数据集的可读性、关联性、可回答性进行了人工评价。实验结果表

明，机器阅读理解数据集的质量是影响模型性能的关键因素之一，同一个模型在不同类型数据集的表现大不相同。对于加入不可回答、知识推理等内容的复杂型数据，目前的藏文机器阅读理解模型并不能取得很好的成绩，此类数据对模型的鲁棒性和理解能力提出了更高的要求。在未来的工作中，我们将继续扩充我们的数据集，并针对藏文更具挑战性的机器阅读理解任务，开展进一步的研究和学习。

致谢

本论文得到了国家自然科学基金项目（61972436）和国家社会科学基金项目（22&ZD035）的资助。

参考文献

Kaustubh Dhole and Christopher D Manning. 2020. Syn-qg: Syntactic and shallow semantic rules for question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 752–765.

Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1342–1352.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46.

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.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *4th International Conference on Learning Representations, ICLR 2016*.

Lynette Hirschman, Marc Light, Eric Breck, and John D Burger. 1999. Deep read: A reading comprehension system. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pages 325–332.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yanghoon Kim, Hwanhee Lee, Joongbo Shin, and Kyomin Jung. 2019. Improving neural question generation using answer separation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6602–6609.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

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.

Shuming Ma, Xu Sun, Wei Li, Sujian Li, Wenjie Li, and Xuancheng Ren. 2018. Query and output: Generating words by querying distributed word representations for paraphrase generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 196–206.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *In Workshop on Cognitive Computing at NIPS*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Xingwu Sun, Jing Liu, Yajuan Lyu, Wei He, Yanjun Ma, and Shi Wang. 2018. Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3930–3939.

Y Sun, S S Liu, C F Chen, Z C Dan, and X B 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.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198.

Liuyin Wang, Zihan Xu, Zibo Lin, Haitao Zheng, and Ying Shen. 2020. Answer-driven deep question generation based on reinforcement learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5159–5170.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Yao Zhao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke. 2018. Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3901–3910.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2018. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing: 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8–12, 2017, Proceedings 6*, pages 662–671. Springer.

Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. 2019. Learning to ask unanswerable questions for machine reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4238–4248.

王丽客, 孙媛, and 刘思思. 2021. 基于多级注意力融合机制的藏文实体关系抽取. 智能科学与技术学报, 3(466-473).