面向机器阅读理解的高质量藏语数据集构建

孙媛[1,2,*] 刘思思[1,2] 陈超凡[1,2] 旦正错[1,2] 赵小兵[1,2]
[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
*Corresponding author: Yuan Sun
tracy.yuan.sun@gmail.com, liusisi.s@qq.com,
chaofanalex@qq.com, 736127388@qq.com, nmzxb_cn@163.com

摘要

机器阅读理解是通过算法让机器根据给定的上下文回答问题，从而测试机器理解自然语言的程度。其中，数据集的构建是机器阅读理解的主要任务。目前，相关算法模型在大多数流行的英语数据集上都取得了显著的成绩，甚至超过了人类的表现。但对于低资源语言，由于缺乏相应的数据集，机器阅读理解研究还处于起步阶段。本文以藏语为例，人工构建了藏语机器阅读理解数据集（TibetanQA），其中包含20,000个问题答案对和1,513篇文章。本数据集的文章均来自云藏网，涵盖了自然、文化和教育等12个领域的知识，问题形式多样且具有一定的难度。另外，该数据集在文章收集、问题构建、答案验证、回答多样性和推理能力等方面，均采用严格的流程以确保数据的质量，同时采用基于语言特征消融输入的验证方法说明了数据集的质量。最后，本文初步探索了三种经典的英语阅读理解模型在TibetanQA数据集上的表现，其结果难以媲美人类，这表明在藏语机器阅读理解任务上还需要更进一步的探索。

关键词： 机器阅读理解 ； 低资源语言 ； 藏语 ； 数据集

Construction of High-quality Tibetan Dataset for Machine Reading Comprehension

Yuan Sun [1,2,*] **Sisi Liu**[1,2] **Chaofan Chen**[1,2] **Zhengcuo Dan**[1,2] **Xiaobing Zhao**[1,2]
[1] Minzu University of China, Beijing 100081
[2] National Language Resource Monitoring & Research Center of Minority Languages
*Corresponding author: Yuan Sun
tracy.yuan.sun@gmail.com, liusisi.s@qq.com,
chaofanalex@qq.com, 736127388@qq.com, nmzxb_cn@163.com

Abstract

Machine reading comprehension asks the machine to answer questions according to the given context through the algorithm, so as to test the degree of understanding natural language text by machines. Among them, the construction of dataset is one of the main tasks in machine reading comprehension. At present, the relevant algorithm models have achieved remarkable results in most popular English datasets, even surpassing the human performance. However, for low-resource languages, due to the lack of corresponding datasets, the research on machine reading comprehension is still in its infancy. Taking Tibetan as an example, this paper constructs Tibetan machine reading comprehension dataset (TibetanQA), which contains 20,000 question answer pairs and 1,513 articles. The articles of dataset are all from Tibetan Yunzang website, which covers 12 topics, such as nature, culture and education. The questions are diverse and difficult. In addition, the data set adopts strict process in the aspects of article collection, question construction, answer verification, answer diversity and reasoning

第二十届中国计算语言学大会论文集，第208页-第218页，呼和浩特，中国，2021年8月13日至15日。

ability to ensure the quality of data, and the verification method based on questions language features shows that the dataset is high quality. Finally, this paper explores the performance of three classic English reading comprehension models on TibetanQA, and the results are not as good as that of humans. This shows that further exploration is needed in the Tibetan machine reading comprehension task.

Keywords: Machine reading comprehension, Low-resource languages, Tibetan, Datasets

1 引言

近年来，机器阅读理解引起了人们的广泛关注，并成为了人工智能研究与应用领域的热点任务之一 (Niu et al., 2020; Gong et al., 2020; Reddy et al., 2020)。它旨在教机器在阅读人类文章后回答出与文章相关的一些问题 (Nguyen et al., 2016)，它需要机器能够理解人类的语言文字。机器阅读理解实际上是一个数据驱动型任务，因此数据集是其技术发展的基础。到目前为止，已经出现了很多大规模的机器阅读理解数据集，比如CNN/Daily Mail (Hermann et al., 2015)、MCTest (Richardson et al., 2013)、CBT (Hill et al., 2015)、RACE (Lai et al., 2017)、SQuAD (Rajpurkar et al., 2016)、DuReader (He et al., 2017)等，这些数据集推动了机器阅读理解的研究，其中以2016年Rajpurkar等人发布的大规模英文机器阅读理解数据集SQuAD为代表，许多学者在这一数据集上提出了自己的方法和模型。根据最新官方数据，截止到2021年3月发布的模型榜单排名在SQuAD2.0的数据集上最高得分为93.183，而人类的得分是89.452[1]。

藏语作为中国少数民族语言之一，由于缺乏公开的藏语机器阅读理解数据集，目前藏语机器阅读理解任务还处于起步阶段。近年来随着互联网的发展，网络上也出现了大量的藏文信息，藏文的信息处理等相关工具也得到了很好的发展 (加羊吉et al., 2014) (色差甲et al., 2019) (夏天赐and 孙媛, 2018) (龙从军et al., 2015)，但如何有效利用这些藏文信息并推动藏文机器阅读理解的发展成为目前藏文信息化建设需要解决的问题之一。

藏语是一种拼音语言，属辅音字母文字型，分辅音字母、元音符号2个部分。其中有30个辅音字母，4个元音字母和5个反写字母（用于拼外来语）。藏文中的语法比较复杂但有很明确的组织形式和动词变化。其中，3个上加字，4个下加字，5个前加字，10个后加字，2个后后加字组合在任意一个基础字理论上便可以写出任何一个藏文词。藏语单词的最小单位是一个音节，一个音节包含一个或最多七个字符，音节默认是用音节之间的标记“.”来分割的。这只是一个上标点。语言文字由一个或多个音节组成，但是由相同的符号”.”分割。辅音簇是用特殊的字母连写而成 (Nuo et al., 2015)。另外，藏文是谓语在后、动词作为核心的语言。在藏文中，有一些特殊的助词，可以清楚地表示句子的语义结构。这些特殊助词主要包括作格助词、属格助词、位格助词和从格助词，例如，作格助词可以表示动作的代理、工具和动作的方式。基于藏语的这些特点，我们可以识别这些格助词来分析藏语机器阅读理解的问题。同时，严格的藏文语法规则要求构建更高水平的藏语机器阅读理解数据集。

为了进一步推动藏语机器阅读理解的发展，同时为了满足相关研究人员对高质量的藏语阅读理解数据集的需求，本文构建了一个藏语机器阅读理解数据集（TibetanQA），并将部分数据公开在`https://tibetanqa.github.io/TibetanQA/index.html`，数据集的示例如表1所示，每个段落下包含多个问题和答案，其中答案来自于文本中。

本文的主要贡献如下：

（1）本文构建了一个藏语机器阅读理解数据集（TibetanQA），其中包含1,513篇藏语文章和20,000个藏语问答对。这些文章均来自云藏网，问答对采用众包的方式人工构建。文章涵盖了12个领域的实体知识，其中包括自然、文化、教育、地理、历史、生活、社会、艺术、技术、人物、科学和体育。

（2）本文采用严格的人工构建流程来保证数据集的质量，尽管目前TibetanQA中问答对的数量不多，但该数据集在文章收集、问题构建、答案验证、回答多样性和推理能力等方面均采用严格的流程以确保数据的质量。另外，数据集中的问答对数量还在不断增加中。

[1]https://rajpurkar.github.io/SQuAD-explorer/

段落	རི་བོང་གི་གཟུགས་གཞིའི་རིང་ཚད་ལ་ལི་སྟེ45ནས་ལི་སྟེ50དང་།ལུས་ཀྱི་སྡིད་ཚད་ལ་སྟེ་ཀྱུ2ནས་སྟེ་ཀྱུ3ཡོད།ནུ་གཤོག་ཤང་ལྟག་ལས་ཀྱང་རིང་བ་རེད།མཆུ་ཞེང་ཆེ་ལ་ཇ་མ་ཀ་ཅོང་ཟུང་།དཔྱར་ཐོག་རྒྱབ་གཞུང་གི་སྐུ་མདོག་རེ་སེར་སྐྱ་ཞེག་ཡིན་ལ།དཔག་ཐོག་རྒྱབ་གི་སྐུ་མདོག་ལ་སྐྱ་ཤས་ཆེས།ཁེར་ཞིང་སོག་ཡིན་པ་དང་ཀྱི་དུ་ཅོང་མཐུག་རེ་བོང་ཐེ་འཇམ་སྤྲིང་ཡོངས་ལ་ཁྱབ་ཡོད། ཆུ་དང་ཐྲོང་བོ་ཡོད་ས་ཡི་ནགས་ཚོད།ཇུ་ཐང་།རི་ལྟང་སོགས་ན་གནས།།རང་ཆུལ་གྱི་མངོ་དཉྩས་མཐོ་སྐང་གི་སྐང་རི་དང་རྫ་ཐང་།ཁྱགས་གཞུང་།ཐང་ཡངས།ནགས་ཚལ།རྫུབས་སོགས་ན་གནས་འདུག	兔子身长45cm-50cm，体重有2kg-3kg。它们的耳朵比腿长，嘴唇很宽尾巴很短。它们背后的毛发在夏天通常为淡黄色而在冬天则大部分为灰色，它们长着一身长毛，毛发松散而且毛根非常厚。兔子遍布世界各地，通常生活在有水和树的森林以及草原和山谷。我国青藏高原的草原、山丘、森林等地都是兔子的栖息地。
问题	རི་བོང་གི་གཟུགས་གཞིའི་རིང་ཚད་ནི་དུ་ཡིན།	兔子身长多少?
答案	རི་བོང་གི་གཟུགས་གཞིའི་རིང་ཚད་ལ་ལི་སྟེ45ནས་ལི་སྟེ50ཡོད།	兔子身长45cm-50cm。
问题	རི་བོང་རང་ཆུལ་གྱི་ས་ཆ་གང་དག་ཏུ་གནས་ཡོད།	兔子分布在我国哪些地方?
答案	རང་ཆུལ་གྱི་མངོ་དཉྩས་མཐོ་སྐང་གི་ཐང་རི་དང་རྫ་ཐང་།ཁྱགས་གཞུང་།ཐང་ཡངས།ནགས་ཚལ།རྫུབས་སོགས་ན་གནས་འདུག	我国青藏高原的草原、山丘、森林等地都是兔子的栖息地。
问题	རི་བོང་གི་ཆུབ་གཞུང་གི་སྐུ་མདོག་ནི་ཅི་ཞེག་ཡིན།	兔子背后的毛发是什么颜色?
答案	དཔྱར་ཐོག་ཆུབ་གཞུང་གི་སྐུ་མདོག་ནི་སེར་སྐྱ་ཞེག་ཡིན་ལ།དཔག་ཐོག་ཆུབ་གཞུང་གི་སྐུ་མདོག་ལ་སྐྱ་ཤས་ཆེ།	兔子背后的毛发在夏天通常为淡黄色而在冬天则大部分为灰色。

表 1. 带有问答对的一个段落举例

（3）本文探索性的以BiDAF、R-Net和QANet三种经典的英语机器阅读理解模型作为TibetanQA数据集上的基线模型，并展开实验，其结果显示模型最好实验结果的F1值比人类表现低21.4%。这表明，在藏语机器阅读理解任务上还需要进行更多的探索。另外，本文采用基于语言特征消融输入的方法进行评估，实验结果表明该数据集对模型的阅读理解能力提出了更高的要求。

2 相关工作

大规模阅读理解数据集是驱动机器阅读理解任务研究发展的重要因素，数据集的质量和规模直接影响到阅读理解模型的理解能力和表现。近年来，出现了大量的机器阅读理解数据集，本文对这些数据集进行调查。按照答案的形式，机器阅读理解数据集可大致分为四个类型：填空型数据集、选择型数据集、篇章片段型数据集和多任务型数据集 (Liu et al., 2019)。表2列举了常见的几种数据集的相关信息。

2.1 填空型数据集

填空型数据集能够将复杂的机器阅读理解问题简化为对一个单词的简单预测，填空型阅读理解的答案是一个单词而非一个句子，机器需要理解上下文的内容来预测段落中丢失的关键词，CNN/Daily Mail (Hermann et al., 2015)语料库中的数据来自美国有线电视新闻网和每日邮报网中的文章，语料库剔除了单篇超过2000个字的文章和问题答案不在原文出现的文章。The children's Book Test (CBT) (Hill et al., 2015) 是经典的阅读理解数据集，它从每个儿童故事中提取20个连续的句子作为文档，第21个句子作为问题，并从中剔除一个实体类单词作为答案，该数据集只关注命名实体识别和普通名词类型的答案预测。

2.2 选择型数据集

选择型阅读理解任务包含一段文章片段和多个问题，每个问题又包含多个选项，要求机器理解给定的文章片段并从给定的答案选项中选出最合适的一个答案。选择型数据集要求能够在一个问题的多个答案候选项中准确地选出答案。MCTest (Richardson et al., 2013)是一组和故事相关的问题集，该数据集的文章来自童话故事，它的问题选项基本为原文中的内容，因此对模型的推理能力要求较低，数据集要求机器能够回答有关虚构故事的多项选择阅读理解问题，直

数据集	文章来源	文章大小	问题大小	语言	种类
CNN/Daily Mail	新闻	300K	1.4M	英文	填空型数据集
CBT	儿童故事	108	688K	英文	填空型数据集
RACE	英语阅读理解	50K	870K	英文	选择型数据集
MCTest	童话故事	500	2K	英文	选择型数据集
SQuAD	维基百科	536	100K	英文	篇章片段型数据集
DuReader	百度搜索和百度知道	1M	200K	中文	多任务型数据集
TibetanQA	云藏网	1.5K	20K	藏文	篇章片段型数据集

表 2. 常见机器阅读理解数据集的大小和文章来源比较

接解决开放域机器理解的高级目标。RACE (Lai et al., 2017)数据集是中国中学生英语阅读理解题目，该数据集规模较大且领域覆盖广泛，题型的正确答案并不一定直接体现在文章中，只能从语义层面深入理解文章，通过分析文章中线索并基于上下文推理，选出正确答案，因此基于该数据集的机器阅读理解模型需要一定的推理能力。

2.3 篇章片段型数据集

篇章片段型阅读理解任务可以描述为：给定一段文章片段，给定一个问题，要求机器根据该问题从文章片段中找到一个连续的片段作为答案。SQuAD (Rajpurkar et al., 2016)是一个大规模的阅读理解数据集。它的文章来自维基百科，并采用众包的方式人工构建问题，该数据集拥有10万以上高质量的问题答案对，由于其高质量和可靠的自动评估，该数据集引起了NLP领域的广泛关注。

2.4 多任务型数据集

多任务型阅读理解任务需要构建高难度的真实世界的数据集，该数据集的问题不限制段落范围，回答一个问题可能需要理解多个段落，并且答案是人为创造的而不是来自文章原文，这就要求机器获得更高的推理能力，从而能真正实现机器阅读理解。DuReader (He et al., 2017)是一个中文阅读理解数据集，该数据集的问题和文章均来自百度搜索和百度知道，答案是人们根据多篇文章推理出来的而不是原始上下文中的片段，DuReader提供了新的问题类型yes、no、和opinion。

目前，英文和中文的阅读理解数据集已经得到了很好的发展，而对于低资源语言的阅读理解数据集则很少有人研究，这严重阻碍了低资源语言的机器阅读理解的发展，为了解决这个问题，我们构建了一个高质量的藏语阅读理解数据集，称为TibetanQA，该数据集的文章来自云藏网，涵盖了多领域的知识，并采用众包的方式人工构建。TibetanQA面向藏语篇章片段型阅读理解任务，数据集中的答案来自文章。

3 构建过程

首先，本文从云藏网上获取了藏文实体的文本，然后对文本信息进行筛选，并采用人工标注的方式构建问答对，最后对问答对的有效性进行人工审核，该工作过程主要包括文章收集、问题构建和答案验证。

3.1 文章收集

为了获取大量的文章，本文利用爬虫技术对云藏网站中的实体知识信息进行爬取，共获取了1,600个实体知识信息文本。其中文本的选取涵盖了广泛的主题，包括自然、文化、教育、地理、历史、生活、社会、艺术、技术、人物、科学、体育共12个领域。此外，本文采用正则表达式对获取到的文章段落中的噪声信息进行处理，删除了图像、表格和网站链接等非文本数据，并丢弃了小于100个音节的段落，最终选取了1,513篇文章。

3.2 问题构建

为了有效地收集问题，我们开发了一个问答收集的Web应用程序如图1所示，并邀请了母语为藏语的学生来使用该应用程序，这些藏族学生从小接受藏语学习，目前为藏学专业研究生，

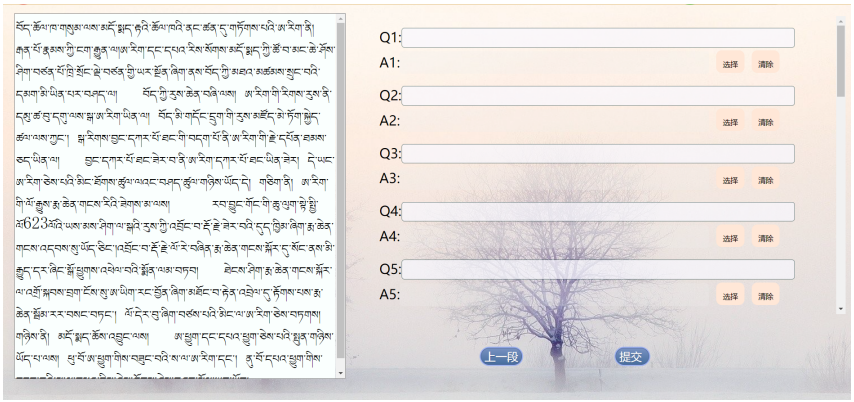

图 1. 用于收集藏语问答对的Web应用程序

方向为语言文学，具有较高的藏文水平。在问题构建的过程中，对于文章中的每一个段落，他们首先需要选择文章中的一段文本作为答案，然后将问题用自己的语言写入输入字段中，学生的任务是提问并回答关于该段落的问题，答案必须为段落中的一部分。当他们完成一篇文章后，系统会自动为其分配下一篇文章。为了构建更具挑战性的语料库，我们对每个学生进行了短期培训，并指导他们如何提供有效和具有挑战性的问题，对于每个学生，我们首先会教其如何进行提问和回答，之后利用少部分数据来对他们进行测试，只有当准确率达到90%的学生才可以进行后面的工作。此外，我们不对问题的形式施加限制，并鼓励他们使用自己的语言来进行提问。

3.3 答案验证

为了进一步提高数据集的质量，我们在获取到初始的数据集后，邀请另一组藏族学生来检查，他们选择有效的问答对，丢弃不完整的答案或问题，剔除语法不正确的问题。最终，我们人工校对出20,000个问题答案对。

4 数据分析

一个高质量的数据集要能够从多个角度对机器阅读理解模型进行准确的评估，因此，本文研究了当前自然语言处理领域中一些流行的机器阅读理解数据集，发现目前流行的数据集主要为英文和中文的，很少有低资源数据集。因此，构建一批有价值和开放的藏语机器阅读理解数据集显得尤为重要。本文重点分析最为权威的篇章片段型机器阅读理解数据集SQuAD，该数据集不仅有超过10万个的问题答案对，而且采用人工构建，可以保证语料库的质量。参考SQuAD数据集的构建方式，我们构建了TibetanQA藏语数据集，为了验证构建的数据集更具挑战性，本文将TibetanQA与SQuAD进行比较，并简要对TibetanQA面临的挑战进行了介绍。

答案类型	TibetanQA	SQuAD
数字	17.6%	10.9%
人名	8.5%	12.9%
地名	7.2%	4.4%
组织机构名	3.5%	-
其他实体	26.5%	21.7%
短语	30.4%	41.2%
日期/时间	6.3%	8.9%

表 3. 不同的答案类型所占的比例

4.1 答案的多样性

本文对TibetanQA中的答案进行了分类，生成答案的音节标签和命名实体识别标签。首先，将数据分为数字答案和非数字答案，之后利用命名实体识别标签将名词短语分为人、地

点、组织机构和其他实体。表3显示不同答案类型所占的比例，其中数字占所有数据的17.6%，日期和时间占6.3%，答案中有30.4%是短语，8.5%是人名，7.2%是地名，3.5%是组织机构名，剩下的26.5%由其他实体和其他类型组成。

疑问词	TibetanQA	SQuAD
什么（ཅི་ཞིག་）	40.4%	61.2%
哪里（གང་）	10.5%	4.8%
谁（སུ་）	8.7%	11.9%
何时（དུས་ནམ་ཞིག་）	11.9%	8.1%
为什么（ཅིའི་ཕྱིར་）	5.4%	1.5%
如何（ཅི་འདྲ་）	14.4%	12.5%
其它（གཞན་དག）	8.7%	-

表 4. TibetanQA和SQuAD中不同类型问句所占的比例

4.2 问题前缀的统计

本文将问句按照疑问词进行分类，通过疑问词将问句分成七种类型：什么（ཅི་ཞིག་）、哪里（གང་）、谁（སུ་）、何时（དུས་ནམ་ཞིག་）、为什么（ཅིའི་ཕྱིར་）、如何（ཅི་འདྲ་）和其它（གཞན་དག），统计结果如表4所示。

从表4中可以看出，疑问词“什么”在两种数据集中的占比均很大，疑问词“哪里”、“谁”、“何时”和“如何”的占比分别为10.5%，8.7%，11.9%和14.4%，这表明TibetanQA中问题类型的分布比较均衡。

4.3 回答问题所需要的推理

为了获得更具有挑战性的数据集，我们在TibetanQA数据集的答案中增加了推理，机器提取正确答案也需要推理能力。我们将所有的问题分成四类：单词匹配、同义词替换、多句推理和模糊问题，类型样例如表5所示。

单词匹配：这类问题是针对段落中的某个关键词（通常为名词）进行提问，即将该关键词用疑问词替换来进行提问，该关键词即为答案，问题中的其余部分的单词均可以直接在文章原文中找到。回答这类问题只需要进行简单的相似算法便可找到答案，不需要任何的推理过程。在表5段落1中，问题中的“蛋白质组成成分”直接在原文中出现，根据一个简单的相似算法便可以确定答案为“氨基酸”。

同义词替换：这类问题在单词匹配的基础上进行了单词的替换，即问题中的关键词不再是原文中的词，而是被与该词意思相近的词所取代，这就在问题和段落之间产生了差异。回答这类问题机器则需要能够识别不同单词之间的相同含义，这可能需要额外的知识。表5段落2中展示了TibetanQA中需要进行同义词替换的数据示例，问题中的“ཨན་གཞི་སྒྲུབ་ ”和段落中的“ཨན་རྒྱང་སྒྲུབ་ ”都是指氨基酸，在这种情况下，我们不能直接将问题和原文进行匹配，需要先进行同义词替换，之后进行单词匹配来获取正确答案。

多句推理：这类问题并不能只根据当前的句子来获得答案，而需要将多个句子组合起来进行简单的推理。表5段落3中展示了TibetanQA中需要进行多句推理类型的数据，其中，我们需要知道代词所指的是什么。在这个例子中，第二个句子中的“它”指的是第一个句子中的“蛋白质”，所以第二个句子可以转换成“蛋白质是人体肌肉的主要成分”，因此，通过指代消解后可以得到答案为“蛋白质”。

模糊问题：这类问题理论上没有标准答案，即不同的人给出的答案可能不同。表5段落4中展示了TibetanQA数据中的模糊性问题，问题为“兔子有什么特点？”，根据段落可知兔子存在许多特点，不同的人会得到不同的答案，因此，在实际的问题中我们会指定一个答案。

5 实验

机器阅读理解数据集的质量直接影响到模型的理解能力，因此需要对构建的数据集进行

段落1	སྤྲེ་དཀར་ཟླས་ཚང་མའི་གྲུབ་ཆ་གཙོ་བོ་ཨན་གཞི་སྣུར་ཡིན།
	蛋白质的所有组成成分都是氨基酸。
类型	单词匹配
问题	སྤྲེ་དཀར་ཟླས་ཚང་མའི་གྲུབ་ཆ་གཙོ་བོ་རེ་ཅི་ཞིག་ཡིན།
	蛋白质的组成成分是什么?
答案	ཨན་གཞི་སྣུར་ཡིན།
	氨基酸。
段落2	ཚན་རིག་པ་ཚོས་ཕྱང་པོ་གསོན་པའི་ནང་ན་ས་**ཨན་གཞི་སྣུར་**རིགས་80རྫུག་ཤེས་གསལ་བྱང་ཡོད་པ་དང་།
	科学家已经在生物体内发现了80多种**氨基酸**。
类型	同义词替换
问题	ཚན་རིག་པ་ཚོས་ཕྱང་པོ་གསོན་པའི་ནང་ན་ས་**ཨན་མང་སྣུར་**རིགས་ག་ཚོད་ཤེས་དགོས་བྱང་ཡོད།
	科学家在生物体内发现了多少种**氨基酸**?
答案	རིགས80རྫུག་
	80多种。
段落3	སྤྲེ་དཀར་ཟླས་ནི་ཕྱང་དང་སྐྱེ་དངོས་ཀྱི་ཕྱང་པོ་གྲུབ་ཆེད་ཀྱི་དངོས་ཟླས་གཙོ་བོ་ཡིན་པ་དང་།མིའི་ལུས་སྐྱེང་གི་ཤ་གནད་ཀྱི་གྲུབ་ཆ་གཙོ་བོ་ཡང་ཡིན།
	蛋白质是人体所有细胞和组织的重要组成部分。它也是人体肌肉的主要组成成分。
类型	多句推理
问题	མིའི་ལུས་སྐྱེང་གི་ཤ་གནད་ཀྱི་གྲུབ་ཆ་གཙོ་བོ་ནི་ཅི་ཞིག་ཡིན།
	人类肌肉的主要组成成分是什么?
答案	སྤྲེ་དཀར་ཟླས།
	蛋白质。
段落4	རི་བོང་གི་གནམ་གཞིའི་རིང་ཚད་ལ་ལི་སྐྱེ45ནས་ལི་སྐྱེ50དང་།ལྟས་ཀྱི་ཕྱེད་ཚད་ལ་སྤྱི་རྒྱ2ནས་སྤྱི་རྒྱ3ཡོད།ནཱ་གཤོག་ཤང་ཟུག་ལས་ཀྱང་རིང་བ་རེད།མཆུ་ཞེང་ཆེ་ལ་ཇ་མ་ཉ་ཅང་ཐུང་།
	兔子身长45cm-50cm，体重2kg-3kg，它的耳朵比腿长，嘴唇宽，尾巴短。
类型	模糊问题
问题	རི་བོང་གི་ལུས་ལ་ཁྱད་ཆོས་ཅི་ཞིག་ཡོད།
	兔子有什么特点。
答案	རི་བོང་གི་གནམ་གཞིའི་རིང་ཚད་ལ་ལི་སྐྱེ45ནས་ལི་སྐྱེ50དང་།ལྟས་ཀྱི་ཕྱེད་ཚད་ལ་སྤྱི་རྒྱ2ནས་སྤྱི་རྒྱ3ཡོད།ནཱ་གཤོག་ཤང་ཟུག་ལས་ཀྱང་རིང་བ་རེད།མཆུ་ཞེང་ཆེ་ལ་ཇ་མ་ཉ་ཅང་ཐུང་།
	兔子身长45cm-50cm，体重2kg-3kg，它的耳朵比腿长，嘴唇宽，尾巴短。

表 5. 各种推理类型举例

评估。本文使用三种经典的英语阅读理解模型R-Net (Wang et al., 2017)、BiDAF (Seo et al., 2016)和QANet (Yu et al., 2018)来测试TibetanQA，并采用一种基于语言特征消融输入的评估方法来直观的评估TibetanQA数据集 (Sugawara et al., 2020)。本文将这三个模型作为基线方法，使用EM和F1来评估模型的准确性。EM是指预测答案和标准答案之间的匹配程度，例如，有m个问题，如果模型能正确回答n个问题，则可以用公式（1）计算EM。

$$EM = \frac{n}{m} \tag{1}$$

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:

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

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

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

其中，$N(TP)$表示预测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。

5.1 在不同模型上的实验

对于传统的数据驱动阅读理解来说，数据集可以使用基于规则的系统 (Riloff and Thelen, 2000)和逻辑回归模型 (Ng et al., 2000)来改进它们的模式匹配基线。尽管这种类型的数据集是真实的和具有挑战性的，但由于数据集太小，无法支持非常有表现力的统计模型。从SQuAD数据集公开发布之后，基于注意力机制的深度学习匹配模型开始大量出现，与传统的基于规则的方法相比，深度学习的方法可以更好地对文本的特征进行表示，能够很大程度上提高模型的学习能力，也就能使机器更好地理解文章内容。

目前，基于数据集的机器阅读理解任务主要采用深度学习的方法进行研究。Seo等人 (Seo et al., 2016)首次引入了“双向注意力机制”的机器阅读理解模型BiDAF，他们在交互层引入段落对问题的注意力和问题对段落的注意力，采用这两个方向的注意力来获得文章和问题之间的表征，他们认为这些注意力可以提取更多的信息。Wang等人 (Wang et al., 2017)首次在机器阅读理解任务中加入自注意力机制并提出了R-Net模型，他们通过计算段落中单词与单词之间的注意力值，学习已经融合了问题信息的段落内部单词之间的权重分布，实验结果证明引入自注意力机制提高了模型的准确率。此外，为了解决RNN在编码过程中会导致训练速度慢的问题，Yu等人 (Yu et al., 2018)将卷积神经网络和自注意力机制结合提出了QANet模型，他们认为提高了训练速度以后可以在同样的时间内训练更多的数据，因此可以提高模型的泛化能力，该模型在SQuAD上取得了更好的成绩。以上三个模型均在SQuAD上取得了不错的成绩，因此本文将BiDAF、R-Net和QANet模型引入到藏语数据集TibetanQA上进行实验。

本文将文章中的段落和问题随机分为训练集和测试集，TibetanQA和SQuAD数据集的统计信息如表6所示。

数据集	训练集		测试集	
	段落	问题	段落	问题
SQuAD	17,007	68,758	1,889	18,841
TibetanQA	5,194	16,000	587	4,000

表 6. 两种数据集的数据统计信息

Model	SQuAD		TibetanQA	
	EM(%)	F1(%)	EM(%)	F1(%)
人类表现	86.831	89.452	87.4	89.2
BiDAF	68.0	77.3	**58.6**	**67.8**
R-Net	71.3	79.7	55.8	63.4
QANet	**73.6**	**82.7**	57.1	66.9

表 7. 不同模型在两种数据集上的实验结果

本文从测试集中随机抽取100个样本，分成10个部分，然后分发给10个不同的藏族学生进行测试，把他们的平均分数作为人类的表现，得到F1值为89.2%。错误匹配的原因主要是藏语中短语的替换和不必要短语的添加或删除，而不是答案的根本分歧。之后，本文分别使用BiDAF、R-Net和QANet模型在TibetanQA上进行了实验，实验结果如表7所示。在SQuAD数据集上，BiDAF模型的EM和F1分别为68%和77.3%。在TibetanQA数据集上，BiDAF模型的EM和F1分别为58.6%和67.8%。R-NET和QANet模型在TibetanQA数据集上的结果也比在SQuAD数据集上的结果要低，主要原因如下：

（1）现有藏文分词工具的错误会传播到下游任务中。

（2）SQuAD的训练集明显多于TibetanQA的训练集。对于低资源语言来说，在小规模数据集上很难获得良好的性能，因此需要机器阅读理解模型来加强模型的理解能力，传统的英语

阅读理解模型不能直接应用到TibetanQA上。因此，该数据集对未来的MRC任务提出了新的挑战。

（3）TibetanQA中的大多数段落长度约为150个词，SQuAD中的大多数段落长度约为100个词，而长文本的段落信息中会存在更多问题无关的信息，因此对模型理解能力的要求也越严格。

5.2 基于语言特征消融输入的评估方法

Saku等人 (Sugawara et al., 2020)提出了基于消融输入的方法来评测机器阅读理解数据集，他们假设输入文本中的某一项特征对应了现实中的一种阅读理解技能，然后通过删除文本中的一些特定语言特征，观察模型结果前后的实验结果来对数据集难度进行验证。他们认为一个数据集在经过某一种消融方法后准确率下降越大，则说明这个数据集对于该技能的要求越严格。反之，如果一个数据集对多数的消融处理都不敏感，则说明它不能有效地评估模型的阅读理解能力。受他们的工作启发，本文针对藏文中词性消融、词顺序、随机单词掩盖以及句子顺序四个角度去评估TibetanQA数据集。

（1）词性消融：将输入的文本信息替换成词对应的词性组成的序列，以便于考察模型是否仅凭词性进行预测。

（2）词顺序：对非答案片段的词顺序进行随机调整，本文以句子为单位针对每个句子中的3个词顺序进行随机替换，目的是考察模型对词序的认知和句子语义构成。

（3）随机单词掩盖：将答案中所在句子一个词随机替换成UNK，以便于考察模型的推理能力。

（4）句子顺序：对段落中句子之间的顺序进行随机的打乱组合，目的是考察机器是否理解句子之间的逻辑。

本文以R-Net模型为基准实验，观察不同的输入对模型效果的影响，计算结果如表8所示。

	F1(%)	增加(%)	EM(%)	增加(%)
R-Net	55.8		63.4	
词性消融	6.2	-49.6	15.8	-47.6
词顺序	35.2	-20.6	42.2	-21.2
随机单词掩盖	26.7	-29.1	35.5	-27.9
句子顺序	31.4	-24.4	43.1	-20.3

表 8. 不同的输入信息对R-Net网络模型的影响

从表8可以看出，四种消融输入对R-Net模型的预测准确率都有所下降，其中以词性消融后的结果最为明显，模型的F1值和EM值分别下降了49.6%和47.6%。这说明模型对词性以外的信息要求较大，单凭词性的特征信息难以获得较高的表现。除此之外，其他三组实验也分别说明数据在词顺序、随机单词掩盖和句子顺序三个方面对模型提出了更高的要求。

6 总结

本文构建了藏语机器阅读理解数据集TibetanQA，其中包含20,000个问题答案对和1,513篇文章。本数据集的文章均来自云藏网，问题答案对采用众包的方式人工构建。TibetanQA为藏语机器阅读理解研究提供数据基础。目前，TibetanQA的基线模型的F1值分别为67.8%、63.4%和66.9%。其性能比人类表现要低21.4%，这表明现有的模型可以在该数据集上可以有更好的改进。接下来，我们将进一步扩展数据集，并鼓励更多的人去探索新的表示模型，以促进低资源语言机器阅读理解的发展。

致谢

本论文得到了国家自然科学基金项目（61972436）资助。

参考文献

Hongyu Gong, Yelong Shen, Dian Yu, Jianshu Chen, and Dong Yu. 2020. Recurrent chunking mechanisms for long-text machine reading comprehensio. *ACL2020*.

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. *Proceedings of the Workshop on Machine Reading for Question Answering*.

Karl Moritz Hermann, Tomáš Kočiskỳ, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 1693–1701.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

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

Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019. Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9(18):3698.

Hwee Tou Ng, Leong Hwee Teo, and Jennifer Lai Pheng Kwan. 2000. A machine learning approach to answering questions for reading comprehension tests. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 124–132.

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 *CoCo@ NIPS*.

Yilin Niu, Fangkai Jiao, Mantong Zhou, Ting Yao, Jingfang Xu, and Minlie Huang. 2020. A self-training method for machine reading comprehension with soft evidence extraction. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Minghua Nuo, Huidan Liu, Congjun Long, and Jian Wu. 2015. Tibetan unknown word identification from news corpora for supporting lexicon-based tibetan word segmentation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 451–457.

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

Revanth Gangi Reddy, Md Arafat Sultan, Efsun Sarioglu Kayi, Rong Zhang, Vittorio Castelli, and Avirup Sil. 2020. Answer span correction in machine reading comprehension. *arXiv preprint arXiv:2011.03435*.

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.

Ellen Riloff and Michael Thelen. 2000. A rule-based question answering system for reading comprehension tests. In *ANLP-NAACL 2000*.

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

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8918–8927.

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.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

加羊吉, 李亚超, 宗成庆, and 于洪志. 2014. 最大熵和条件随机场模型相融合的藏文人名识别. 中文信息学报, 28(1):107–112.

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.

色差甲, 贡保才让, and 才让加. 2019. 藏文音节拼写检查的cnn 模型. 中文信息学报, 33(1):111–117.

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.