藏文文本校对评测集构建

三毛措[1,2,3,5]　**才智杰**[1,2,3,4,5]　**道吉扎西**[1,2,3,5]

1. 青海师范大学计算机学院，青海西宁 810016；

2. 藏文信息处理教育部重点实验室，青海西宁 810008；

3. 青海省藏文信息处理与机器翻译重点实验室，青海西宁 810008；

4. 西南民族大学计算机科学与技术学院，四川成都 610041；

5. 藏语智能信息处理及应用国家重点实验室，青海西宁 810008

2627996852@qq.com　Czjqhsd@163.com　1336786645@qq.com

摘要

文本校对评测集是拼写检查研究的基础，包括传统文本校对评测集和标准文本校对评测集。传统文本校对评测集是对正确的数据集通过主观经验人工伪造而得到的评测集，是一种常用的文本校对评测方式，但也存在诸多的缺陷。标准文本校对评测集是通过选择研究对象获取可信度强的真实数据集而得到的评测集。本文在分析英、汉文文本校对评测集构建方法的基础上，结合藏文的特点研究了藏文文本校对评测集的构建方法，构建了用于评价藏文文本校对性能的标准文本校对评测集，并统计分析了评测集中的错误类型及分布，以此验证本文构建的标准文本校对评测集的有效性和可用性。

关键词： 自然语言处理 ； 藏文 ； 文本校对 ； 评测集

Construction of Tibetan text proofreading evaluation set

SAN Maocuo[1,2,3,5]　***CAI Zhi-jie***[1,2,3,4,5]　***DAO Jizhaxi***[1,2,3,5]

1.College of Computer Science and Technology, Qinghai Normal University, Qinghai Xining 810016;

2.Key Laboratory of Tibetan Informati on Processing,Ministry of Education,Qinghai Xining 810008;

3.Tibetan Information Processing and Machine Translation Key

Laboratory of Qinghai Province,Qinghai Xining 810008;

4.School of Computer Science and Technology, Southwest Minzu University, Sichuan Chengdu 610041;

5.The State Key Laboratory of Tibetan Intelligent Information

Processing and Application, Qinghai Xining 810008

2627996852@qq.com　Czjqhsd@163.com　1336786645@qq.com

Abstract

Text proofreading evaluation set is the basis of spelling research, including traditional text proofreading evaluation set and standard text proofreading evaluation set. The traditional text proofreading evaluation set is a kind of evaluation set obtained by artificial forgery of the correct data set through subjective experience. It is a common text proofreading evaluation method, but it also has many defects. Standard text proofreading evaluation set is an evaluation set obtained by selecting research objects to obtain reliable real data sets. Based on the analysis of the construction methods of English and Chinese text proofreading evaluation set, combined with the characteristics of Tibetan, this paper studies the construction methods of Tibetan text proofreading evaluation set, constructs a standard text proofreading evaluation set for evaluating the performance of Tibetan text proofreading, and analyzes the error types and distribution in the evaluation set, so as to verify the effectiveness of the standard text proofreading evaluation set constructed in this paper Availability and usability.

Keywords: Natural Language Processing , Tibetan , Text proofreading , Evaluation set

第二十届中国计算语言学大会论文集，第229页-第237页，呼和浩特，中国，2021年8月13日至15日。

1 引言

随着自然语言处理研究的不断深入，技术方法的评测成为自然语言处理的研究内容之一，评测集是评测的基础数据，有了合理的评测集才能准确的评测方法的有效性。文本校对评测集是用于评价文本校对效果的数据集，可分为传统文本校对评测集和标准文本校对评测集。传统文本校对评测集是对正确的数据集通过主观经验人工伪造而得到的评测集，标准文本校对评测集是通过选择研究对象获取可信度强的真实数据集而得到的评测集。在没有构建标准评测集的情况下，通常使用传统文本校对评测集进行评测，由于传统评测集是人工伪造的数据集，不能覆盖文本校对的各种类型，因此学者们开始研究标准评测集的构建。

藏文文本校对研究刚刚起步，目前还没有用于评价藏文文本校对性能的标准评测集。随着大数据时代的不断推进，藏文电子语料也与日俱增，但网上爬取的语料几乎都是经过相关机构的审核，基本无错误。本文结合藏文的这种特点并分析英、汉文文本校对评测集构建方法的基础上，研究了藏文文本校对评测集的构建方法，构建了用于评价藏文文本校对性能的标准评测集，并统计分析了评测集中的错误类型及分布。

2 研究现状

文本校对评测是文本校对性能分析的基础，评测的目的是验证模型的性能，以比较各种文本校对技术的优劣。然而，评测体系需要客观公正，不受主观感觉影响。自 2014 年起，学者们开始了文本校对评测集的建设工作。英文文本校对评测集建设方面，ACL 自然语言学习特别兴趣小组（SIGNLL）组织了 CoNLL 年度会议，是专门用于探讨自然语言处理技术方法，而他们 2014 年的任务是探讨文本校对评测集方法。此小组在研究英语语法错误检测技术时以标准的方式构建了 50 篇论文的英文文本校对评测集 (Ng H T, 2014)，是用于评测英语语法错误检测，该评测集数据采集对象是 25 名非英语国家大学的学生，其构建方式是根据给出的两个提示每人写两篇论文，评测集的具体信息如表 1 所示。在 CoNLL-2014 语法错误检测任务综述中，17 个小组采用不同的文本校对方法在相同的英文文本校对评测集上验证其任务的性能，学者们希望在这样的平台上能够挖掘出更先进的英语语法错误检测技术。

Test data	Size
Essays	50
Sentences	1381
Word tokens	29207

表 1: 英文文本校对评测集信息表

汉文文本校对评测集建设方面，2015 年自然语言处理技术研讨会 NLP-TEA 与中国语法错误检测（CGED）共同为汉文文本校对工具的开发和实施提供了一个论坛。他们在研究汉语语法错误检测时以标准的方式构建了 1000 个评测句的汉语文本校对评测集 (Lung-Hao Lee, 2015)，是用于评测汉语语法错误检测，该评测集数据采集内容是台湾地区的 TOCFL 机考作文，其构建方式是以汉语为母语的人手工标注语法错误，并提供相应的纠正，然而以开放测试的方式进行评估，促进了汉语文本校对技术的发展。2017 年由台湾国立大学、计算语言与中文处理协会主办、亚洲自然语言处理联合协会（AFNLP）承办的第八届国际自然语言处理联席会议 IJCNLP

2017 的共同任务也是汉语语法错误检测，他们为此以标准的方式构建了用于汉语文本校对的评测集 (Rao, 2017)，是用于评测汉语语法错误检测，该评测集的数据采集内容是《汉语水平考试》的写作部分，其构建方式与上面方法一致，表 2 显示了评测集中错误类型的分布。2018 年国际自然语言处理与中文计算会议 NLPCC2018 的共同任务中他们在研究汉语语法纠错时从北京大学汉语学习语料库中抽取 2000 个句子以标准的方式构建了一个汉语文本校对评测集 (Zhao Y, 2018)，其目的是评测汉语语法纠错。

Error Type	Redundant words	Missing words	Word selection errors	Word ordering errors
Number	1062	1274	2155	385
Percentage	21.78%	26.13%	44.20%	7.90%

表 2: IJCNLP 2017 语法错误检测评测集中错误类型的分布

国内少数民族语言文字的文本校对评测集构建研究才刚刚起步，藏文文本校对评测集构建研究也处于探索阶段。目前，学者们普遍采用爬虫技术抓取网络上现有的语料，进而通过增加噪声数据的方法获取训练集或者评测集（传统评测集）。2018 年才智杰等 (2018) 在研究向量模型的藏文非真字自动拼写检查时采用传统的方式构建了规模为 11.7 万的藏文字评测集，2019 年色差甲等 (2019) 在研究 CNN 藏文音节拼写检查时采用传统方式构建了人工伪造的音节数据集，2020 年华旦扎西等 (2020) 在研究 TC-LSTM 的藏文词拼写检查时也采用传统的方式主观构建了规模为 400 句的藏文词评测集。综上所述，我们可以看出英文和汉文文本校对评测集建设方面已比较成熟，同时也推动了英文和汉文的自然语言处理发展。但藏文文本校对评测集构建方面还未见标准评测集的相关文献报道，制约了藏文文本校对技术的发展。

3 藏文文本校对评测集构建

3.1 藏文文本校对评测集构建方案

数据采集是评测集构建的第一步。数据采集方式有多种多样，通常采用问卷调查、做实验、查阅资料、下载公开数据集、人工伪造数据、爬虫以及现场采集等方式。目前藏文文本校对技术研究时学者们采用人工伪造数据构建评测集（传统评测集），传统评测集存在错误类型覆盖率不全、可靠性较低，不能准确反映文本校对的真实情况。现场采集的数据能反映出问题的真实性，具有直观性、科学性、真实性以及通用性等优点。因此，我们借鉴英文和汉文的文本校对评测集的建立过程，采用现场采集的方式进行了评测数据采集，以此设计了藏文文本校对评测集的构建方案。藏文文本校对评测集的构建方案如图 1 所示。

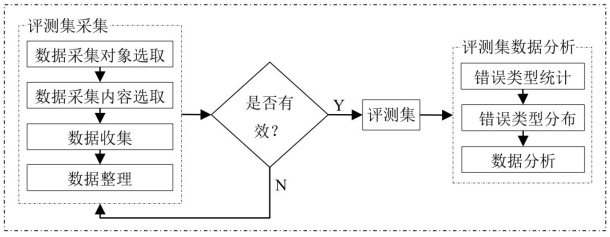


图 1: 藏文文本校对评测集构建方案

藏文文本校对评测集构建方案由评测集采集和评测集数据分析两部分组成，评测集采集包括数据采集对象选取、数据采集内容选取、数据收集以及数据整理等四部分，评测集数据分析包

括错误类型统计、错误类型分布及数据分析等三部分。构建藏文文本校对评测集时首先要选择对自己研究问题相符合的数据采集对象，进而选择数据采集内容，其次进行现场采集数据样本，再次对采集的数据样本进行数据整理，然后判断数据是否有效，最后对构建的评测集进行数据分析。数据整理是将采集到的数据进行规范化，并将其数字化。根据数据集中是否出现拼写错误判断数据集的有效性，若有拼写错误则将其归入评测集，否则重新采集。数据分析部分通过统计评测集中的错误类型和分析错误类型的分布情况，验证评测集构建的合理性和有效性。

3.2 藏文文本校对评测集构建

根据藏文文本校对评测集构建方案，可以按以下步骤建立藏文文本校对评测集。

第一步：数据采集对象选取

我们将拉加草原学校作为这次课题研究数据采集的对象。该学校位于青海省果洛藏族自治州玛沁县拉加镇，这所学校设有 9 个年级，14 个教学班，开设的课程与其他中小学的设课内容基本一致，所使用的教材均为教育部标准教材。此外，将辩论的课程和思想也被运用在了日常教学中。该校的学生都从小学一年级开始学习藏语，都是母语学习者，因而藏语水平总体上比其他语言文字成绩普遍都较高，学生们不仅来自青海各个州县，还有甘肃、四川等地区的学生，有助于捕捉到各个地方的在藏文文本中出现的各种拼写错误类型。学校设有 9 个年级（一年级至九年级），其中一年级到三年级的藏语学习时长较短，还未掌握很多的藏文知识点，藏语水平较低，因此本研究以四年级至九年级共六个年级作为研究对象进行现场数据采集。数据采集对象信息如表 3 所示。

序号	数据采集对象	人数	序号	数据采集对象	人数
1	四年级学生	42	2	五年级学生	79
3	六年级学生	88	4	七年级学生	85
5	八年级学生	43	6	九年级学生	85

表 3: 数据采集对象信息表

表 3 列举的信息可作为本课题研究对象的主要原因有以下四方面，第一：由于该学校的学生均为母语使用者，为数据的有效性提供了基础的保障；第二：由于该学校收集到的数据具有多元化，使得数据类型应较全面及覆盖率较高，并与下游任务藏文文本校对的实验内容相吻合；第三：由于每个年级每位学生的藏语水平各不相同，使得我们采集到的数据具有很大的研究价值；第四：由于我们课题组将采取现场采集的方式采集数据，使得我们采集到的数据具有真实性。综上，我们的数据采集对象满足数据的有效性、多样性、价值性，真伪性等四大特性，符合作为本次研究的数据采集对象。

第二步：数据采集内容选取

选取研究内容时，我们考虑学生的藏语学习时长的长短，计划从三年级到八年级的上下册教材书中各选一篇课文，共 12 篇课文。为了提高数据的质量和数据采集的效率，低年级的课文普遍都很短，我们选择其中常用词较多篇长较长的课文作为数据采集的内容，高年级的课文普遍都很长，我们选择其中常用词较多篇长较短的课文作为数据采集的内容。因此本文选取的数据采集内容的覆盖性较全面，其中的常用词也较普遍，符合作为本课题的数据采集内容。藏文文本校对评测集数据采集内容的选取信息见表 4 所示。为了获取学生在已学课文的情况下所犯的真实性的拼写错误数据，我们将四年级的学生作为三年级上下册教材书两篇课文的数据采集的

对象，将五年级的学生作为四年级上下册教材书两篇课文的数据采集的对象，以此类推。

序号	数据采集对象	数据采集内容	序号	数据采集对象	数据采集内容
1	四年级学生	ངག་ཚལ་ཅ་བའི་རྩ་བྱ།(骄傲的孔雀) ཚོགས་པའི་དཔེ་ཆྭག(晨读)	2	五年级学生	མཆེ་བ།(麻雀) ཉི་མ་རྒྱ་མཚོའི་ངོས་ནས་འཆར་བ།(太阳从海面升起)
3	六年级学生	ཐྲག་བཟོའི་བུ།(鞋匠的儿子) ཁྱུ་ཀུ།(燕子)	4	七年级学生	ས་བོན་གྱི་སྟོབས་ཤུགས།(种子的力量) བྱང་ཆུབ་ཤིང་།(菩提树)
5	八年级学生	དཔེ་ཆན་ལ་གུས་པར་བྱ་བ་དང་བག་ཡོད་སྤྱད་རྣམས།(尊敬老师，感恩师长)ངའི་ཀི་དཔེ་ཆན།(我的老师)	6	九年级学生	ངའི་ནུ་ཐན་འདི་ལ་དགའ།(我爱青青小草) ཁང་རེས་རྟ་ཁ།(千里马)

表 4: 评测集数据采集内容选取信息表

第三步：数据的收集

由于初三年级需要准备中考和缺少人手等的种种原因，我们最终实际获取到的数据只有六篇课文的内容，共 232 名学生的研究样本，数据信息见表 5 所示。数据采集的方式是现场采集的方法，即老师到每个班听写相应的课文，对每个数据采集对象采取一致的数据采集方法，保证采集数据样本的有效性、可比性、可靠性及研究价值，使得采集的数据具有普遍性和代表性。表 5 中数据采集参与人数和最终收集的数据样本数一致。

序号	1	2	3	4	5	6	共计
数据采集对象	四年级学生	五年级一班学生	五年级二班学生	六年级学生	七年级学生	八年级学生	——
数据采集内容	ཚོགས་པའི་དཔེ་ཆྭག	མཆེ་བ།	ཉི་མ་རྒྱ་མཚོའི་ངོས་ནས་འཆར་བ།	ཐྲག་བཟོའི་བུ།	ས་བོན་གྱི་སྟོབས་ཤུགས།	དཔེ་ཆན་ལ་གུས་པར་བྱ་བ་དང་བག་ཡོད་སྤྱད་རྣམས།	——
数据采集人数	40 人	36 人	33 人	36 人	42 人	45 人	232 人

表 5: 实际评测集数据采集信息表

第四步：数据的整理

评测数据的代表性决定了最终建立的藏文文本校对评测集的可靠性和可行性。完成第三步的数据收集后，我们对收集到的纸质版数据样本进行了数据整理。数据整理包括对数据样本进行编号，例如四年级 40 名学生的数据样本编号依次为 4-1、4-2、…、4-40，五年级一班 36 名学生的数据样本编号依次为 5(1)-1、5(1)-2、…、5(1)-36，五年级二班 33 名学生的数据样本编号依次为 5(2)-1、5(2)-2、…、5(2)-33，以此类推。由藏语为母语的人对样本进行人工检查拼写错误并做了错误注释；将 6 个数据采集内容（正确的数据内容）进行电子化（文档）并以.TXT 文本格式分别放置已新建的 6 个文件中；在每个文件中按每个数据采集参与人数复制粘贴相应的文档并对文档进行编号（纸质版样本编号数 = 文档复制数 = 数据采集参与人数 = 文档编号数），然而它们之间是一对一的关系；按照每份纸质版样本中注释的错误将在对应电子文档中的正确的字改成错误的字，获取最终的藏文文本校对评测集，藏文文本校对评测集信息如表 6 所示，表中的数据大小是指评测集文档数的总大小，例如序号 1 对应的数据大小 200KB 是指四年级 40 名学生的评测集文档数的总大小有 200KB，以此类推。

序号	数据采集内容	数据采集人数	评测集文档数	字数/平均	句数/平均	数据大小/KB
1	ཚོགས་པའི་དུས་ཐོག	40 人	40	364	31	200KB
2	མཆེས་པ།	36 人	36	387	40	180KB
3	ཉེ་མ་རྒྱ་མཚོའི་ངོས་ནས་འཆར་བ།	33 人	33	405	40	165KB
4	ཤྭ་བཟའི་བྱ།	36 人	36	579	52	252KB
5	ས་བོན་གྱི་སྒོ་བས་ཤུགས།	42 人	42	661	54	336KB
6	དགེ་ཏན་ལ་གུས་པར་བྱ་བ་དང་བགའ་དིན་རྣམས་སུ་དྲན་པ།	45 人	45	484	45	270KB
共计	–	232 人	232	2880	262	1403KB

表 6: 藏文文本校对评测集信息分布表

4 藏文文本校对评测集数据分析

根据文献 (San Maocuo, 2021) 中归纳的藏文文本真字型错误类型，我们课题组首先识别了上面构建的藏文文本校对评测集中的拼写错误类型，其次对错误类型的分布进行了统计，最后根据统计表对数据进行分析。

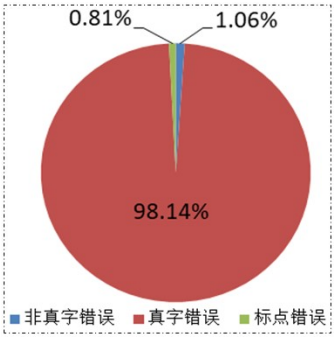
为了进一步弄清评测集中错误类型的分布，我们做了更详细的统计和分析，得出了以下评测集中错误类型分布情况的统计结果，整个藏文文本校对评测集中的错误类型的分布统计如表 7 所示。表 7 中全集是指整个 232 个评测集文档之和。为了更直观的了解评测集中的拼写错误类型的分布，根据表 7 中统计的数据本文画出了对应的饼图，如图 2(a) 所示。

错误类型	全集（232）		
	数量	占比	排次
非真字错误	171	1.06%	2
真字错误	15905	**98.14%**	1
标点错误	131	0.81%	3

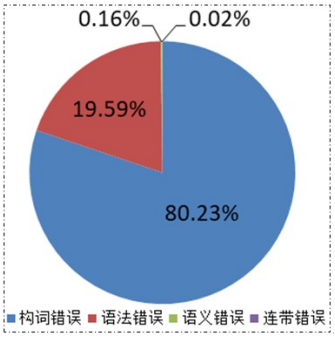
表 7: 全样本错误类型的分布统计表

藏文文本拼写错误类型分为非真字错误、真字错误和标点错误三大类，由表 7 可知，此三大错误类型的占比分别依次为 1.06%、98.14%、0.81%，其中藏文文本中的错误类型主要集中在真字错误，例如将词“རྒྱ་མཚོ”（大海）写成“རྒྱལ་མཚོ”属于真字拼写错误，这应该是受各地藏区方言的影响而导致拼写出错，在我们日常中发音“རྒྱ་མཚོ”（大海）这个词时往往将第一个字“རྒྱ”读成“རྒྱལ”，而在听写时受此影响导致拼写错误。其次是非真字错误，例如将“སྟོབས་ཤུགས”（力量）写成“སྟོབས་ཤུག”是属于藏文音节缩略的非真字错误，虽然在一些手书和手抄本中这种写法很常见，但是这种书写方式并非正式规范，而且也不符合藏文字的拼写规则，学生们仍然习惯用缩略的方式拼写词汇，从而导致非真字错误。最后是标点错误，它在藏文文本中占的比例最小，例如将“ཡིན་རྟང་”（但是）写成“ཡིན་ཏ་བང་”属于标点冗余错误。根据本课题构建的藏文文本校对评测集中的错误类型的统计结果可知，评测集中的拼写错误类型和与我们日常拼写中存在的错误类型较吻合，说明我们构建的评测集是非常合理的，并且错误类型的分布情况也是合理的。由于真词错误是目前藏文文本校对领域中最主要的一部分，下面单独对此进行了详细统计和分析，如表 8 所示。为了更直观的了解评测集中的真字错误类型，根据表 8 中统计的数据本文画出对应的饼图，如图 2(b) 所示。

真字拼写错误是词、语法和语义层面的拼写错误，而且有些词的错误导致其后的虚词添加错误。基于上述问题，藏文真字型错误类型包括构词错误、语法错误、语义错误及连带错误四小



(a) 全样本错误类型的分布图



(b) 真字错误类型的分布图

图 2: 错误类型分布图

错误类型		全集（232）		
		数量	占比	排次
真字错误	构词错误	12761	**80.23%**	1
	语法错误	3116	19.59%	2
	语义错误	25	0.16%	3
	连带错误	3	0.02%	4

表 8: 真字错误类型的分布统计表

类，四个错误类型的占比分别依次为 80.23%、19.59%、0.16%，0.02%。其中构词错误占了很大一部分，造成此错误的最大原因可能是学生的学习方法有所欠缺，对字或词的掌握存在死记硬背的现象，未了解字或词的构成规则，不理解字或词的真正含义，使得造成对字或词的部分记忆缺失，出现大量的字或词的拼写错误，说明学生对词汇基础不牢，尤其是在动词、名词的使用上犯了很多错误，例如将“ཞྭ་བཟོ་བ”（鞋匠）错写成“ཞྭ་གཟོ་བ”。其次是语法错误，语法错误主要是不自由虚词的添加错误和动词时态错误，例如句子“ངས་ཁྱོད་དུས་ནས་ཨ་ཕས་བྱང་ནས་ཞྭ་བཟོའི་ལག་རྩལ་སྦྱངས་ཡོད།”（我从小跟我爸学修鞋技术）中“ཕས”（应是“ཕའི”）违反了属格助词的添加规则，属于语法错误。构词错误和语法错误能够最直接地反映学生对藏语词汇基础的掌握情况。由于本评测集是以课文听写的方式进行采集，而不是学生自由发挥，因此其中存在的语义错误和连带错误相对较少，本文不将对此进行重点说明。

为了能更清楚的观察每篇课文的评测集中每个错误类型的分布情况，本文又详细统计了这些数据信息，具体信息见表 9 所示。表 9 中数量是指对于一篇课文（共 6 篇课文）的每个评测集文档中出现该错误类型的总数。根据表 9 中统计的数据本文得出了每篇课文的藏文文本校对评测集中错误类型的分布图，见图 3 和图 4。

错误类型		课文 1（40）			课文 2（36）			课文 3（33）		
		数量	占比	排次	数量	占比	排次	数量	占比	排次
非真字错误		15	1.70%	3	41	1.19%	3	48	1.35%	3
真字错误	构词错误	656	**74.21%**	1	2721	**79.17%**	1	2934	**82.60%**	1
	语法错误	211	23.87%	2	649	18.88%	2	547	15.40%	2
	语义错误	0	0.00%	5	5	0.15%	5	7	0.20%	5
	连带错误	2	0.23%	4	1	0.03%	6	0	0.20%	6
标点错误		0	0.00%	5	20	0.58%	4	16	0.45%	
错误类型		**课文 4（36）**			**课文 5（42）**			**课文 6（45）**		
		数量	占比	排次	数量	占比	排次	数量	占比	排次
非真字错误		18	0.88%	3	26	0.82%	3	23	0.74%	4
真字错误	构词错误	1589	**77.29%**	1	2291	**72.41%**	1	2570	**82.53%**	1
	语法错误	434	21.11%	2	826	26.11%	2	449	14.42%	2
	语义错误	1	0.05%	5	2	0.06%	5	10	0.32%	5
	连带错误	0	0.00%	6	0	0.00%	6	0	0.00%	6
标点错误		14	0.68%	4	19	0.60%	4	62	1.99%	3

表 9: 各篇评测集中错误类型的分布统计表

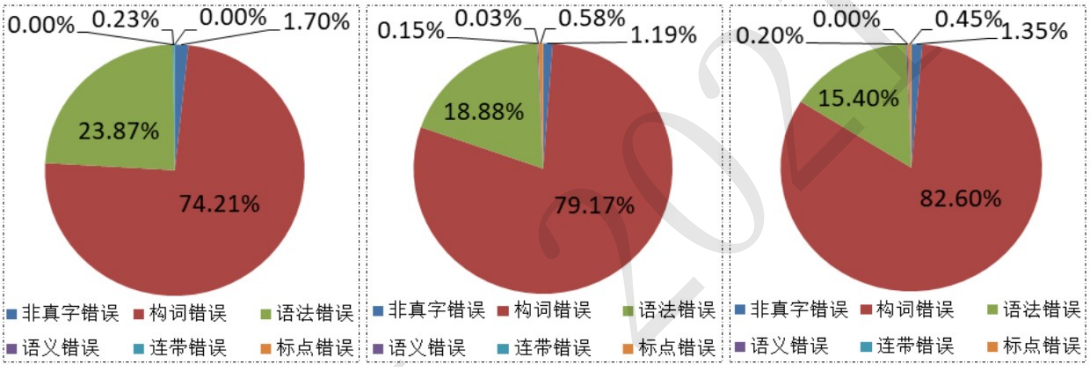


图 3: 课文 1 至 3 评测集中错误类型分布情况

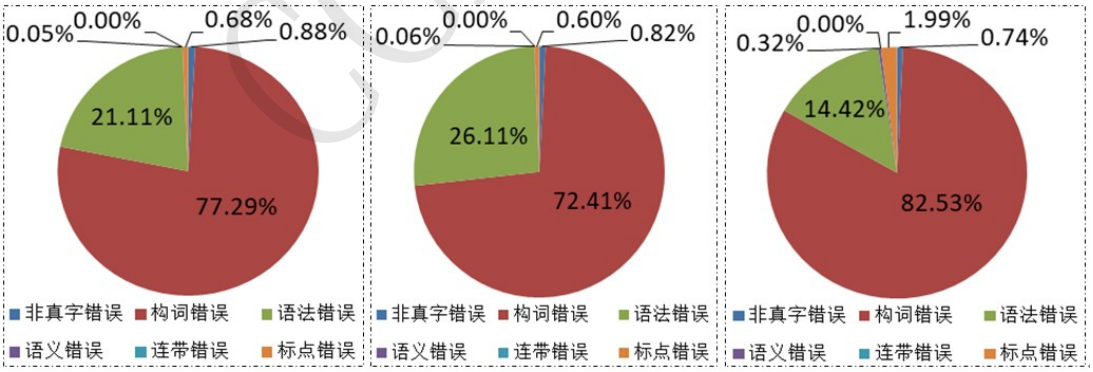


图 4: 课文 4 至 6 评测集中错误类型分布情况

表 9、图 3 和图 4 可知，各篇评测集中构词错误占的比例最大，占比在 72.41%~82.6% 的区间，其次是语法错误，占比在 14.42%~26.11% 的区间，再次是非真字错误，占比在 0.74%~1.70% 的区间。至于语义错误、连带错误和标点错误是由于本文的研究内容和采集方式的选择，因而出现的频次很少，但也符合实际情况，因此本文构建的藏文文本校对评测集是合理的。

5 结论

本文通过分析英文和汉文文本校对评测集构建方法，设计了藏文文本校对评测集构建方案，根据此方案构建了藏文文本校对评测集，并统计分析了藏文文本校对评测集中存在的错误类型。统计数据结果表明，从藏文文本错误类型大类层面来说，错误类型主要集中在真字错误，占比为98.14%。藏文真字错误类型又分为构词错误、语法错误、语义错误及连带错误四小类，其中构词错误占的比例最高，占比为 80.23%。这就决定了我们在研究下一任务藏文文本校对方法时应该注重哪一类错误类型，如何选择任务的侧重点，为藏文文本校对技术研究奠定了基础。本文数据采集的合理性、可操作性及统计分析结果确保了我们构建的评测集的有效性。今后将在已构建的藏文文本校对评测集的基础上，研究藏文真字的文本校对方法，进一步完善自动文本校对技术。

致谢

本项工作得到了国家自然科学基金资助项目 (61866032,61966031), 青海省科技厅资助项目(2019-SF-129), “长江学者和创新团队发展计划” 创新团队资助项目（IRT1068）, 青海省重点实验室项目（2013-Z-Y17、2014-Z-Y32、2015-Z-Y03）, 藏文信息处理与机器翻译重点实验室（2013-Y-17、2020-ZJ-Y05）, 青海师范大学大学生创新创业训练计划项目（qhnucxcy2020070）资助。

参考文献

Ng H T, Wu S M, Briscoe T, et al. 2014. *The CoNLL-2014 Shared Task on Grammatical Error Correction*. Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task.

Lung-Hao Lee, Liang-Chih Yu, Li-Ping Chang. 2015. *Overview of the NLP-TEA 2015 Shared Task for Chinese Grammatical Error Diagnosis*. Workshop on natural language processing techniques for educational applications; Annual meeting of the Association for Computational Linguistics; International joint conference on natural language processing.

Rao, G, Zhang, B, Xun, E, Lee, L. 2017. *IJCNLP-2017 Task 1: Chinese grammatical error diagnosis*. Proceedings of the IJCNLP 2017, Shared Tasks, pp. 1–8. Asian Federation of Natural Language Processing, Taipei (2017).

Zhao Y, Jiang N, Sun W, et al. 2018. *Overview of the NLPCC 2018 Shared Task: Grammatical Error Correction*. CCF International Conference on Natural Language Processing and Chinese Computing. Springer, Cham, 2018:439-445.

才智杰, 孙茂松, 才让卓玛. 2018. 一种基于向量模型的藏文字拼写检查方法. 中文信息学报,2018,32(09):47-55.

色差甲, 贡保才让, 才让加. 2019. 藏文音节拼写检查的 *CNN* 模型. 中文信息学报,2019,33(01):111-117.

华旦扎西, 才智杰, 班玛宝. 2020. 一种基于 *TC-LSTM* 的藏文词拼写检查方法. 中文信息学报,2020,34(05):50-55.

Barukčić I., San Maocuo, Cai Zhijie, Cai Rangzhuoma, Dao Jizhaxi. 2021. *Analysis on types of spelling errors in true Tibetan characters*. MATEC Web of Conferences,2021,336.