智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

一种基于形容词知识库的藏文文本数据增强方法

仁青吉,才智杰

引用本文:
仁青吉, 才智杰. 一种基于形容词知识库的藏文文本数据增强方法[J]. 智能系统学报, 2026, 21(2): 519-528.
REN Qingji, CAI Zhijie. A method for enhancing Tibetan text data based on adjective knowledge base[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(2): 519-528.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202503033

DOI: 10.11992/tis.202503033
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20260202.1508.003

一种基于形容词知识库的藏文文本数据增强方法

仁青吉[1,2], 才智杰[1,2]
（1. 青海师范大学 计算机学院, 青海 西宁 810016; 2. 藏语智能信息处理及应用国家重点实验室, 青海 西宁 810008）

摘　要: 基于深度学习的自然语言处理领域中, 数据集质量和规模直接影响模型的性能。数据增强作为扩展和丰富数据集的有效手段, 是自然语言处理中不可或缺的重要技术之一。文章针对藏文数据资源匮乏的问题, 结合实际语料分析了藏文形容词的语义、情感以及修饰对象等特征, 将藏文形容词按语义特征及修饰对象分为描述性质、状态、数量、感官和感受等 5 大类 46 小类, 通过提取藏文形容词和形容词修饰对象的特征构建了藏文形容词知识库和形容词修饰对象近义词表, 提出了一种基于形容词知识库的藏文文本数据增强方法。该方法通过匹配形容词的类型、音节数等特征替换形容词, 同时匹配形容词修饰对象的句式结构, 将形容词修饰对象用近义词表中对应的词替换。实验结果表明, 该方法能够显著增加藏文文本数据量, 在小学一年级至六年级藏文课本句子集上的总增长率达 990.22%; 在下游任务中也有良好表现, 预训练模型为 RoBERTa、TiBERT、TBERT 和 CINO 时 SimCSE 模型的相关系数分别提升了 8.78、3.17、0.61 和 1.33 百分点, 文本分类任务中准确率、召回率和 F1 值分别提升了 5.97、9.51 和 9.31 百分点。

关键词: 自然语言处理; 低资源语言; 藏文; 形容词; 知识库; 数据增强; 修饰对象; 句式结构
中图分类号: TP391　　**文献标志码**: A　　**文章编号**: 1673−4785(2026)02−0519−10

中文引用格式: 仁青吉, 才智杰. 一种基于形容词知识库的藏文文本数据增强方法 [J]. 智能系统学报, 2026, 21(2): 519–528.
英文引用格式: REN Qingji, CAI Zhijie. A method for enhancing Tibetan text data based on adjective knowledge base[J]. CAAI transactions on intelligent systems, 2026, 21(2): 519–528.

A method for enhancing Tibetan text data based on adjective knowledge base

REN Qingji[1,2], CAI Zhijie[1,2]
(1. College of Computer Science and Technology, Qinghai Normal University, Xining 810016, China; 2. The State Key Laboratory of Tibetan Intelligent Information Processing and Application, Xining 810008, China)

Abstract: In the field of natural language processing based on deep learning, the quality and scale of datasets directly impact model performance. Data augmentation is an essential technique in natural language processing, serving as an effective means to expand and enrich datasets. This paper addresses the issue of Tibetan data resource scarcity by analyzing the semantic, emotional, and modifying object features of Tibetan adjectives based on actual corpora. Tibetan adjectives are categorized into five main categories—descriptive properties, states, quantities, sensations, and feelings—which include a total of forty-six subcategories. By extracting the features of Tibetan adjectives and their modifying objects, a knowledge base for Tibetan adjectives and a synonym table for modifying objects were constructed. We propose a data augmentation method based on this knowledge base, which replaces adjectives by matching their types and syllable counts, while also substituting modifying objects with corresponding synonyms based on their syntactic structures. Experimental results indicate that this method can significantly increase the volume of Tibetan text data, achieving a total growth rate of 990.22% on sentence sets derived from Tibetan language textbooks for grades one through six.It also shows strong performance in downstream tasks. When RoBERTa, TiBERT, TBERT, and CINO are used as the pretrained models, the correlation coefficient of the SimCSE model increases by 8.78, 3.17, 0.61, and 1.33 percentage points, respectively. In the text classification task, accuracy, recall, and F1 score are improved by 5.97, 9.51, and 9.31 percentage points, respectively.

Keywords: natural language processing; low-resource languages; tibetan language; adjectives; knowledge base; data augmentation; modified object; sentence structure

近年来, 藏语自然语言处理逐渐成为研究藏语语言技术的一个重要方向, 资源匮乏是制约其发展的重要因素之一。数据增强是解决藏语资源

收稿日期: 2025−03−24.　**网络出版日期**: 2026−02−02.
基金项目: 国家自然科学基金项目 (61866032, 61966031); 青海省科技厅项目 (2019-SF-129); 藏文信息处理教育部重点实验室项目 (2020-ZJ-Y05).
通信作者: 才智杰. E-mail: Czjqhsd@163.com.

匮乏的有效性方法，词替换[1]是一种常见的数据增强方法，包括两种不同的词替换方式。第一种是用任意词或某类词替换句子中的词，对替换后得到的句子结构和逻辑不做严格要求，生成的数据含有一定的噪声，主要用于扩充负例数据；第二种是用特定的词替换句子中的某类词，既要求替换后得到的句子结构正确还要求符合逻辑，主要用于扩充数据。例如，句子“སློབ་མས་ཏོག་ཅོ་གློག”（学生读书）中，将词“དཔེ་ཆ”（书）替换为“ཙོག་ཅོ”（桌子）后得到的句子“སློབ་མས་ཏོག་ཅོ་གློག”（学生读桌子）虽然语法结构正确，但不符合逻辑，属于第一种词替换方式；句子“མཛེས་ལྗོག་གི་རྩྭ་བང་།”（美丽的草原）中，将词“རྩྭ་བང”替换为“མེ་ཏོག”（花朵）后得到的句子“མཛེས་ལྗོག་གི་མེ་ཏོག”（美丽的花朵）语法结构正确而且符合逻辑，属于第二种词替换方式。数据匮乏是目前藏语自然语言处理存在的主要问题，目前主要采用第二种方式增强数据。形容词[2]在藏文句子中扮演着重要的角色，用来修饰主、谓、宾，使用频率非常高，其运用也非常灵活。因此以第二种词替换方式替换形容词应该是藏语数据增强的一种有效途径。本文在分析藏语语料中形容词的语义、情感以及修饰对象等特征的基础上，通过构建藏文形容词知识库，基于形容词知识库研究了第二种词替换方式下的藏文文本数据增强方法。

1　研究现状

数据增强技术[3]是一种利用已有数据的特征和结构，对原始数据进行变换和扩充生成新的数据样本，以增加数据量和数据多样性的方法，是解决数据资源匮乏和多样性的有效途径。

自然语言处理主要采用词汇替换[4-5]（同义词替换和任意词替换）、增加噪声[6-7]、回译[8-10]和句子重组[11-12]等数据增强方法，任意词汇替换、增加噪声法对原始数据使用删除、插入等添加噪声[13]，以扩充负例数据，提升模型的鲁棒性；同义词替换、回译、句子重组法通过替换原始数据中某些词，用其同义词替换或对原始数据翻译为其他语言，再翻译为源语言或调换原始数据中的词序，用于扩充数据，以增加数据量。Zhang 等[14]、Wei 等[15]基于同义词库替换同义词，以扩充汉语数据量。Coulombe[16]采用替换同义词、缩写动词、转换否定和情态动词等数据增强方法，扩充了汉语语料数据量，有效提升了文本分类性能。Fadaee 等[17]通过选择语料中频率低于阈值的低频词，利用长短期记忆网络（long short-term memory, LSTM）模型将特定上下文中低词替换为常见词，以增加数据的多样性，从而提升神经机器翻译的质量。张蓉等[18]针对方面级情感分析领域中标签数据较难获取的问题，进行句子级相邻词、领域级同类词和词向量级同义词替换的数据增强策略，在 SemEval2014 Task4 Sub Task2 上进行实验，结果表明该数据增强方法是有效的。尤丛丛等[19]利用小规模的平行语料，首先通过对单语词向量的学习，获得一端语言低频词的同义词列表；然后对低频词进行同义词替换，再利用语言模型对替换后的句子进行筛选；最后将筛选后的句子与另一端语言中的句子进行匹配的增强方法，以获得汉越平行数据集，实验结果表明该方法在汉越翻译任务上取得了很好的效果。

藏文是一种低资源语言，藏语自然语言处理中面临的最大困难是数据资源匮乏的问题。汪超[20]在研究藏汉机器翻译方法时，采用低频词用同义词替换、藏汉平行语料库回译等方法有效扩充语料库规模，提高了翻译质量。色差甲等[21]采用混淆子集随机替换音节和根据上下文信息对语义相似音节进行替换的数据增强方法，在藏文 La 格分类、藏文命名实体识别和文本分类任务取得了较好的效果。

形容词作为藏文的重要修饰成分，其丰富的语义和高频率使其成为数据增强的理想对象。鉴于此，本文通过分析藏文形容词的特点，提出了一种基于形容词知识库的藏文文本数据增强方法。

2　基于形容词知识库的藏文文本数据增强

2.1　基于形容词知识库的藏文文本数据增强模型

如果采用形容词用同义词替换，而不考虑上下文和语法规则增强数据，可能导致语义不连贯和逻辑混乱。例如，句子“མོས་ལྱ་བ་ཡག་པོ་ཞེག་གྱོན་ཡོད།”（她穿着一件漂亮的衣服）中的形容词“ཡག་པོ”（漂亮）用形容词“མཛེས་ལྗོག”（美丽）、“མཛེས”（美）替换，得到的句子“མོས་ལྱ་བ་མཛེས་ལྗོག་ཞེག་གྱོན་”（她穿着一件美丽的衣服）词语搭配不当，得到的句子“མོས་ལྱ་བ་མཛེས་ཞེག་གྱོན་ཡོད།”（她穿着一件美的衣服）语法结构不正确。出现上述问题的原因有：一是没有对形容词按照语义进行细致分类，导致生成的句子逻辑不合理，并在语义上不连贯，使得句子难以理解或者表达的意思不准确且不符合自然语言的习惯用法。二是若将单音节形容词修饰的对象用双音节或多音节形容词修饰，导致生成的句子语法结构不合理，并且韵律、节奏不连贯和语言节奏不协调，使得句子不流畅。为解决以上问题，本文建立了形容词知

识库，通过对形容词按语义特征及修饰对象分类，并在形容词知识库中增添更加具体的形容词类型特征项，确保替换后的句子语义连贯、逻辑合理且节奏协调，进而设计了基于形容词知识库的藏文文本数据增强模型，如图 1 所示。

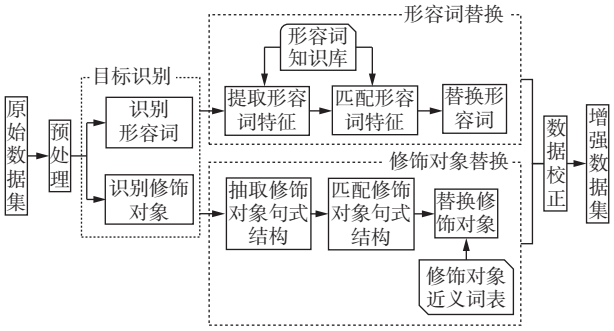

图 1　基于形容词知识库的藏文文本数据增强模型

Fig. 1　Tibetan text data augmentation model based on an adjective knowledge base

基于形容词知识库的藏文文本数据增强模型由预处理模块、目标识别模块、形容词替换模块、修饰对象替换模块和数据校正模块组成。预处理模块的功能是清洗原始数据集、分词和词性标注等操作；目标识别模块的功能是识别预处理后数据集中的形容词及其修饰对象；形容词替换模块依据已构建好的形容词知识库提取形容词特征，并与知识库中形容词的特征进行匹配，利用匹配成功的形容词替换原句中的形容词；修饰对象替换模块的功能是抽取已识别的形容词及修饰对象的句式结构，与句式结构表中句式匹配的修饰对象用近义词表中的词替换；数据校正模块用于校正增强文本中的语法错误等问题。

2.2　基于形容词知识库的藏文文本数据增强

2.2.1　藏文形容词知识库构建

2.2.1.1　藏文形容词知识库

形容词知识库是基于形容词知识库的藏文文本数据增强的重要组成部分，形容词分类是构建形容词知识库的基础。传统语言学[22-24]按照语义将藏文形容词分为颜色、拟形、拟声、尺度、性质、状态、触觉、味觉、数量和动作等，这些分类不能有效捕捉形容词之间的上下文特征和语义差异，不能满足数据增强的需求。为了建立面向数据增强的形容词知识库，本文在传统语言学形容词分类的基础上，结合实际语料分析将藏文形容词按语义特征及修饰对象划分为描述性质、状态、数量、感官和感受等 5 大类，并将 5 大类细分为 46 小类。基于语义特征及修饰对象的藏文形容词分类见表 1。

表 1　基于语义特征及修饰对象的藏文形容词分类

Table 1　Classification of Tibetan adjectives based on semantic features and modification targets

序号	类型	标记	描述对象	描述特征	小类											
					类型名	标记	类型名	标记	类型名	标记	类型名	标记	类型名	标记	类型名	标记
1	性质	XZ	人、事物	特征、品质或性质	厚薄	BH	稀稠	XC	粗细	CX	大小	DX	形态	XT	高低	GD
					面积	MJ	年龄	NL	软硬	RY	锐钝	RD	深浅	SQ	松紧	SJ
					速度	SD	人品	RP	曲直	QZ	形状	XZ	颜色	YS	长短	CD
					真假	ZJ	智力	ZL	凹凸	AT	重轻	ZQ	距离	JL		
2	状态	ZT	人、事物	状态或情况	新旧	XJ	贫富	PF	干净	GJ	声音	SY	清晰	QX	心理	XL
					生理	SL	关系	GX	拟声	NS	情感	QG	表情	BQ	动作	DZ
					主次	ZC	行为	XW								
3	数量	SL	人、事物	数量、范围	多少	DS	范围	FW								
4	感官	GG	人	感官知觉	嗅觉	XJ	触觉	CJ	味觉	WJ	听觉	TJ	视觉	SJ		
5	感受	GS	人	感受、观点或评价	人评	RP	物评	WP								

形容词的修饰对象是形容词的重要知识，是构建形容词知识库的要素之一。藏文形容词主要修饰名词和动词[25]，为了在知识库中描述形容词修饰对象的特征，形容词知识库需要添加修饰对象属性项。另外，形容词的音节数对语法结构的正确性具有重要的作用，因此形容词知识库还需要添加音节属性项。

根据以上分析，替换形容词时需要综合考虑形容词类型、形容词所含音节数和形容词修饰的对象，因此本文基于小学和初高中教材语料构建了包含形容词类型、音节数和修饰对象的用于形容词替换的知识库，共有 860 条形容词。形容词

知识库结构描述如下：

Typedef Struct mold //类型定义
{string b_{road}; //大类
string s_{ub}; //小类}
Typedef Struct modify //对象定义
{bool n; //能否修饰名词
bool v; //能否修饰动词}
Typedef Struct TiAdj_DB //形容词知识库
{string a_{dj}; //存储形容词
Struct mold t_{ype}; //形容词类型
int s_{yll}; //音节数
struct modify o_{bject}; //修饰对象}

藏文形容词知识库 TiAdj_DB 中 a_{dj} 表示形容词，t_{ype} 表示形容词类型，s_{yll} 表示形容词音节数，o_{bject} 表示形容词修饰对象，n 表示是否可以修饰名词，当 n 值为 T 时表示该形容词可以修饰名词，n 值为 F 时表示该形容词不能修饰名词，v 表示是否可以修饰动词，当 v 值为 T 时表示该形容词可以修饰动词，v 值为 F 时表示该形容词不能修饰动词，b_{road} 表示形容词大类，s_{ub} 表示形容词小类。形容词知识库实例见表 2。

表 2　形容词知识库实例

Table 2　Examples from the adjective knowledge base

a_{dj}	t_{ype}		s_{yll}	o_{bject}	
	b_{road}	s_{ub}		n	v
དམར	XZ	YS	1	T	F
མགྱོགས་པོ	XZ	SD	2	F	T

如表 2 所示，形容词“དམར”为“性质（XZ）”大类中的“颜色（YS）”小类，音节数为 1，能修饰名词（n 为 T，v 为 F）。又如形容词“མགྱོགས་པོ”为“性质（XZ）”大类中的“速度（SD）”小类，音节数为 2，能修饰动词（n 为 F，v 为 T）。

2.2.1.2　形容词修饰对象近义词表构建

本文通过观察发现，形容词的修饰对象用其近义词替换能有效扩充数据规模。例如，句子“དེའི་དེབ་ལེགས་པོ་ཉོས།”（买好的书籍）中，形容词“ལེགས་པོ”（好）的修饰对象“དེའི་དེབ”（书籍）用近义词“དེའི་ཚ”（书本）和“སྒྲིག་དེབ”（读本）替换可以得到语义符合逻辑且语法结构正确的句子“དེའི་ཚ་ལེགས་པོ་ཉོས།”（买好的书本）和“སྒྲིག་དེབ་ལེགས་པོ་ཉོས། ”（买好的读本），从而能增强文本数据。藏文形容词及其修饰对象（名词和动词）间有 8 种句式结构，在这 8 种句式结构中形容词的修饰对象可以用该修饰对象的近义词替换。形容词与修饰对象间的句式结构见表 3。

表 3　形容词修饰名词与动词句式结构

Table 3　Sentence structures of adjective modification for nouns and verbs

序号	类型	修饰对象	句式结构
1	名词的定语	名词	$n+a$ $a+g_z+n$ $n+g_z+a$ $n+g_l+a$ $n+\overline{n}+a$
2	动词的状语	动词	$a+g_l+v$ $a+g_x+v$
3	动词的宾语	动词	$a+v$

形容词修饰名词与动词的句式结构表定义为 SyntaxDB，包括类型、修饰对象和句式结构等数据项。形容词修饰名词时作定语，有 5 种句式结构：“$n+a$”表示名词之后为形容词；“$a+g_z+n$”表示形容词在前名词在后，形容词与名词之间有属格助词 g_z；“$n+g_z+a$”表示名词在前形容词在后，形容词与名词之间有属格助词 g_z；“$n+g_l+a$”表示名词在前形容词在后，形容词与名词之间有位格助词 g_l；“$n+\overline{n}+a$”表示名词与形容词之间存在非名词性词。形容词修饰动词时作状语和宾语，有 3 种句式结构：“$a+g_l+v$”表示形容词在前动词在后，形容词与动词之间有位格助词 g_l；“$a+g_x+v$”表示形容词在前动词在后，形容词与动词之间有作格助词 g_x；“$a+v$”表示形容词在前动词在后。

本文从小学和初高中的原始语料中提取形容词的修饰对象，添加修饰对象近义词，构建了 444 组共 1 134 个形容词修饰对象的近义词表。

2.2.2　基于形容词知识库的藏文文本数据增强方法

在构建藏文形容词知识库及形容词修饰对象近义词表的基础上，通过匹配形容词的类型、音节数等特征替换形容词，同时根据形容词修饰对象句式结构的类型将形容词修饰对象用近义词表中对应的词替换。

基于形容词知识库的藏文文本数据增强时，从包含形容词的藏文句子中识别形容词和形容词修饰对象，判断包含形容词的句子是否符合形容词修饰名词与动词的 8 种句式结构。如果符合则用形容词知识库中与该形容词属性值（大类、小类、音节数和修饰对象）相同的形容词替换，并用修饰对象近义词库中的近义词依次替换修饰对象；如果形容词句不符合形容词修饰名词与动词的 8 种句式结构则只替换形容词。例如句子“བོད་རྒེལ་འཚུབ་ཀྱིས་ཚོ་ཏོག་བཟེགས།”（农民急忙地收割粮食）中，形容词“རྒེལ་འཚུབ”（急忙）的修饰对象为“ཞིང་བ”（农民）（属于“$n+\overline{n}+a$”型句式结构），形容词“རྒེལ་འཚུབ”（急

忙）属于“性质（XZ）”大类中的“速度（SD）”小类，音节数为 2，修饰对象可以为名词与动词（n 为 T，v 为 T），从知识库中匹配到属性值相同的形容词“འཚབ་འཚུབ”（急躁）、“རྩེལ་སྐྱུར”（着急）和“རྩེལ་བ”（忙碌）等，替换形容词得到句子“ཞིང་བས་རྩེལ་འཚུབ་ཀྱིས་ལོ་ཏོག་ལེགས།”（农民急忙地收割粮食）、“ཞིང་བས་འཚབ་འཚུབ་ཀྱིས་ལོ་ཏོག་ལེགས།”（农民急躁地收割粮食）、“ཞིང་བས་རྩེལ་སྐྱུར་གྱིས་ལོ་ཏོག་ལེགས།”（农民着急地收割粮食），“ཞིང་བས་རྩེལ་བས་ལོ་ཏོག་ལེགས།”（农民忙碌地收割粮食）；在此基础上用形容词修饰对象近义词表中的近义词“ཞིང་པས་པ”（农民）、“ཤོ་ནམ་པ”（农夫）和“རྫོ་འདེབས་མཁན”（耕作者）等替换修饰对象，得到句子“ཞིང་བས་རྩེལ་འཚུབ་ཀྱིས་ལོ་ཏོག་ལེགས།”（农民急忙地收割粮食）、“ཞིང་པས་པས་རྩེལ་འཚུབ་ཀྱིས་ལོ་ཏོག་ལེགས།”（农民急忙地收割粮食）、“ཤོ་ནམ་པས་རྩེལ་འཚུབ་ཀྱིས་ལོ་ཏོག་ལེགས།”（农夫急忙地收割粮食）、“རྫོ་འདེབས་མཁན་གྱིས་རྩེལ་འཚུབ་ཀྱིས་ལོ་ཏོག་ལེགས།”（耕作者急忙地收割粮食）、“ཞིང་བས་འཚབ་འཚུབ་ཀྱིས་ལོ་ཏོག་ལེགས།”（农民急躁地收割粮食）、“ཞིང་པས་པས་འཚབ་འཚུབ་ཀྱིས་ལོ་ཏོག་ལེགས།”（农民急躁地收割粮食）、“ཤོ་ནམ་པས་འཚབ་འཚུབ་ཀྱིས་ལོ་ཏོག་ལེགས།”（农夫急躁地收割粮食）、“རྫོ་འདེབས་མཁན་གྱིས་འཚབ་འཚུབ་ཀྱིས་ལོ་ཏོག་ལེགས།”（耕作者急躁地收割粮食）、“ཞིང་བས་རྩེལ་སྐྱུར་གྱིས་ལོ་ཏོག་ལེགས།”（农民着急地收割粮食）、“ཞིང་པས་པས་རྩེལ་སྐྱུར་གྱིས་ལོ་ཏོག་ལེགས།”（农民着急地收割粮食）、“ཤོ་ནམ་པས་རྩེལ་སྐྱུར་གྱིས་ལོ་ཏོག་ལེགས།”（农夫着急地收割粮食）、“རྫོ་འདེབས་མཁན་གྱིས་རྩེལ་སྐྱུར་གྱིས་ལོ་ཏོག་ལེགས།”（耕作者着急地收割粮食）、“ཞིང་བས་རྩེལ་བས་ལོ་ཏོག་ལེགས།”（农民着急地收割粮食）、“ཞིང་པས་པས་རྩེལ་བས་ལོ་ཏོག་ལེགས།”（农民忙碌地收割粮食）、“ཤོ་ནམ་པས་རྩེལ་བས་ལོ་ཏོག་ལེགས།”（农夫忙碌地收割粮食）、“རྫོ་འདེབས་མཁན་གྱིས་རྩེལ་བས་ལོ་ཏོག་ལེགས།”（耕作者忙碌地收割粮食）。具体数据增强过程如图 2 所示。基于形容词知识库和修饰对象近义词表的文本数据增强算法见附录。

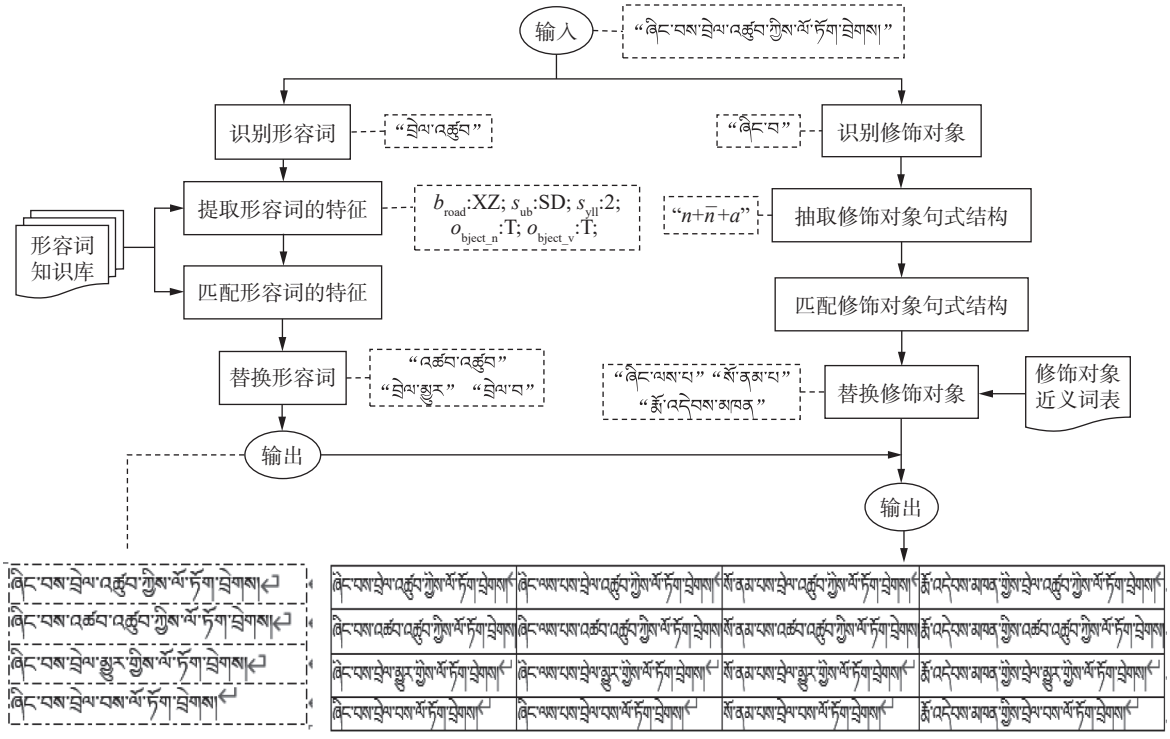

图 2　数据增强过程

Fig. 2　Data augmentation process

3　实验及数据分析

为了验证基于形容词知识库的藏文文本数据增强方法的有效性，本文从数据增强的数量和对下游任务的作用两方面进行分析，设计了 4 组实验。实验选用了数据集 TPSTC、TSTS-B 和 TNCC[26]，其中数据集 TPSTC (Tibetan primary school textbook corpus) 是本文构建的小学一年级至六年级藏文课本中的句子集，包含 11 048 个句子；数据集 TSTS-B(Tibetan semantic textual similarity benchmark) 是用机器翻译系统翻译汉文句向量表示通用公开数据集 STS-B[27]，并人工校对得到的用于藏文句向量表示的数据集，包含 5 231 个句子；数据集 TNCC(Tibetan news classification corpus) 是由复旦大学构建的藏文新闻分类数据集，用于文本分类任务研究。

3.1　数据增强数量分析

实验 1　数据增强数量增长率实验

实验 1 以数据集 TPSTC 为对象，按年级观测了基于名词近义词替换、形容词近义词替换和形容词知识库的藏文文本数据增强方法增强的数量，用 Repl_Noun 表示替换名词的近义词，Noun_GR

表示名词近义词替换的增长率，Repl_Adj 表示替换形容词的近义词，Adj-GR 表示形容词近义词替换的增长率，Repl_Obj 表示替换形容词修饰对象，Repl_Adj_Obj 表示同时替换形容词和形容词修饰对象，Adj_Obj_GR 表示替换形容词和形容词修饰对象的增长率。数据集 TPSTC 中，一年级教材语料中共有 98 个句子，用 Repl_Noun 增强后得到 182 句，Noun_GR 为 85.71%，用 Repl_Adj 增强后得到 2 024 句，Adj-GR 为 1 965.31%，用 Repl_ Obj 增强后得到 154 句，用 Repl_Adj_Obj 增强后得到 2 178 句，Adj_Obj_GR 为 2 122.45%；二年级教材语料中共有 1 050 个句子，用 Repl_Noun 增强后得到 1 962 句，Noun_GR 为 86.86%，用 Repl_Adj 增强后得到 13 054 句，Adj-GR 为 1 143.24%，用 Repl_ Obj 增强后得到 1 312 句，用 Repl_Adj_Obj 增强后得到 14 366 句，Adj_Obj_GR 为 1 268.19%；三年级教材语料中共有 1 790 个句子，用 Repl_Noun 增强后得到 3 407 句，Noun_GR 为 90.33%，用 Repl_Adj 增强后得到 23 843 句，Adj-GR 为 1 232.01%，用 Repl_Obj 增强后得到 2 225 句，用 Repl_Adj_Obj 增强后得到 26 068 句，Adj_Obj_GR 为 1 356.31%；四年级教材语料中共有 2 520 个句子，用 Repl_Noun 增强后得到 4 519 句，Noun_GR 为 79.33%，用 Repl_Adj 增强后得到 23 324 句，Adj-GR 为 825.56%，用 Repl_Obj 增强后得到 3 036 句，用 Repl_Adj_Obj 增强后得到 26 360 句，Adj_Obj_GR 为 946.03%；五年级教材语料中共有 2 420 个句子，用 Repl_Noun 增强后得到 4 647 句，Noun_GR 为 92.02%，用 Repl_Adj 增强后得到 20 604 句，Adj-GR 为 751.40%，用 Repl_Obj 增强后得到 2 822 句，用 Repl_Adj_Obj 增强后得到 23 426 句，Adj_Obj_GR 为 868.01%；六年级教材语料中共有 3 170 个句子，用 Repl_Noun 增强后得到 6 127 句，Noun_GR 为 93.28%，用 Repl_Adj 增强后得到 24 396 句，Adj-GR 为 669.59%，用 Repl_Obj 增强后得到 3 653 句，用 Repl_Adj_Obj 增强后得到 28 049 句，Adj_Obj_GR 为 784.83%。数据集 TPSTC 数据增强情况见表 4 及图 3。

表 4　数据集 TPSTC 数据增强情况

Table 4　Data augmentation statistics of the TPSTC dataset

年级	Original	Repl_Noun	Noun_GR	Repl_Adj	Adj_GR	Repl_Obj	Repl_Adj_Obj	Adj_Obj_GR
一年级	98	182	85.71	2 024	1 965.31	154	2 178	**2 122.45**
二年级	1 050	1 962	86.86	13 054	1 143.24	1 312	14 366	**1 268.19**
三年级	1 790	3 407	90.33	23 843	1 232.01	2 225	26 068	**1 356.31**
四年级	2 520	4 519	79.33	23 324	825.56	3 036	26 360	**946.03**
五年级	2 420	4 647	92.02	20 604	751.40	2 822	23 426	**868.01**
六年级	3 170	6 127	93.28	24 396	669.59	3 653	28 049	**784.83**
TPSTC	11 048	20 844	88.67	107 245	870.72	13 202	120 447	**990.22**

注：加粗表示最优结果。

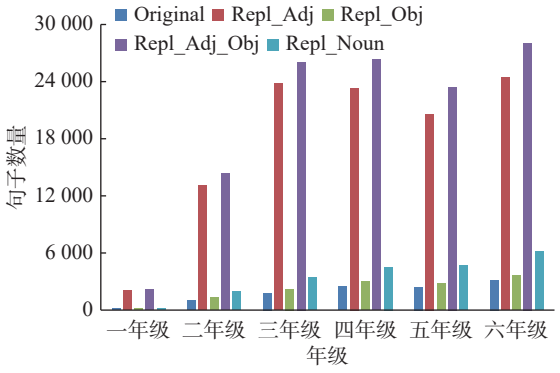

图 3　数据集 TPSTC 数据增强数量分布

Fig. 3　Distribution of data augmentation quantity in the TPSTC dataset

由表 4 及图 3 可见，1）数据增强的效果与原始数据集的规模密切相关，随着原始数据集规模的扩大，数据增强的效果变得更加显著，表明在较大的数据集上本文数据增强方法能够显著提高句子数量，从而更有效地扩展数据集的容量和多样性。2）4 种增强方法表现出明显的差异，Repl_Adj_Obj 增强方法的增长率最高，相较于 Repl_Noun 增长率提升了 901.55%，原因在于形容词和形容词修饰对象是文本中语义变化丰富的成分，生成的样本多样性更高；而 Repl_Noun 和 Repl_Obj 可替换的同义词或变体有限，导致样本量扩充幅度较小。3）同时使用形容词替换法和修饰对象替换法能够充分发挥两种方法的优势，增强了文本描述的丰富性，使数据增强效果最佳，数据增长率为近 10 倍。

实验 2　增强数据质量分析

为了评估数据增强生成的数据质量，本文从 TPSTC 数据集上用 Repl_Noun 和 Repl_Adj_Obj 两

种方法生成的数据中按类型随机抽取了 1% 的语料，从句子的结构正确率、语义连贯性和逻辑合理性 3 个维度进行了人工评估，评估数据见表 5。

表 5　数据增强质量分析
Table 5　Analysis of data augmentation quality　%

方法	类型	数据量	语法结构正确率	语义及逻辑合理性	均值
Repl_Noun	—	208	97.1	66.3	81.70
Repl_Adj_Obj	性质	378	100.0	83.3	91.65
	状态	376	99.2	85.6	92.40
	数量	85	98.8	89.4	94.10
	感官	68	100.0	75.0	87.50
	感受	295	99.3	84.1	91.70

由表 5 中实验数据可见，Repl_Noun 增强方法的语法结构正确率为 97.1%，Repl_Adj_Obj 增强方法的语法结构正确率平均为 99.5%，表明两种增强方法对语法结构的干扰很小，这是因为无论是替换名词的近义词还是形容词都不改变词性和词位置，即替换词保持在原有的词位置上替换原有的词性，很少会破坏句子的语法结构，因此语法结构整体正确率较高。Repl_Noun 增强方法的语义及逻辑合理性为 66.3%，Repl_Adj_Obj 增强方法的语义及逻辑合理性平均为 83.5%，较 Repl_Noun 高出 17.2 百分点，表明了 Repl_Adj_Obj 增强方法生成的数据在语义及逻辑合理性方面质量较高。这是由于 Repl_Noun 易导致语义场冲突，影响逻辑合理性，例如句子“ཉིམ་པས་ཐིམ་པ་ནི་སྡོད་ལའི་རྣམ་པ་ཡིན།”（小孩画的是秋天的景象）中的名词“རྟེམ་པ་”替换为“བཙོ་ཐུ་”，得到的句子“ཉིམ་པས་དྲིས་པ་ནི་སྡོད་ལའི་བཙོ་ཐུ་ཡིན།”（小孩画的是秋天的样式）语义及逻辑不合理；Repl_Adj_Obj 基于形容词知识库，替换后的句子结构正确，语义连贯，符合逻辑，例如句子“མཚན་མོའི་མཁའ་དབྱིངས་སུ་སྐར་མ་ཤིན་ཏུ་མང་ཞ་”（夜空中有很多星星）中的形容词“མང”替换为“མོད”，得到的句子“མཚན་མོའི་མཁའ་དབྱིངས་སུ་སྐར་མ་ཤིན་ཏུ་ མོད།”（夜空中有很多星星）语义连贯且符合逻辑。

3.2　下游任务中的效果分析

实验 3　数据增强对句向量表示的贡献实验

本实验选用 Gao 等[28] 于 2021 年在 EMNLP（Proceedings of the Conference on Empirical Methods in Natural Language Processing）上提出的句向量表示性能优良的无监督表示模型 SimCSE（similarity-based contrastive learning of sentence embeddings），以数据集 TPSTC 为训练集，以数据集 TSTS-B 为测试集，以评估句向量表示性能的斯皮尔曼相关系数 ρ 作为评估标准。由于 SimCSE 模型的第二个模块采用 BERT（bidirectional encoder representations from Transformers）风格的句嵌入编码架构，因此需要选择兼容 BERT 接口的预训练模型。目前公开的藏文预训练模型主要包括 4 种：1）青海师范大学发布的 RoBERTa[29]（a robustly optimized BERT pretraining approach）模型；2）青海师范大学和兰州大学联合发布的 TBERT[30]（Tibetan bidirectional encoder representations from Transformers）；3）中央民族大学发布的 TiBERT[31]（Tibetan bidirectional encoder representations from Transformers）；4）哈工大讯飞联合实验室发布的 CINO[32]（Chinese minority pre-trained language model）模型。这些模型不仅能够独立地表示藏文句子，还可以与 SimCSE 模型进行组合使用。本实验将这 4 个藏文预训练模型分别与 SimCSE 模型相结合，句向量的池化使用了 last-avg 方法。模型参数设置和实验数据信息见表 6 和表 7。

表 6　模型参数设置
Table 6　Model parameter settings

预训练模型	Batch_Size	Learning rate	Dropout rate	Epoch
RoBERTa	32	10^{-5}	0.1	1
TiBERT	32	10^{-5}	0.1	1
TBERT	32	10^{-5}	0.1	1
CINO	32	10^{-5}	0.1	1

表 7　数据增强在句向量表示中的实验数据
Table 7　Experimental results of data augmentation on sentence embeddings　%

预训练模型	句向量模型	增强前ρ	增强后ρ
RoBERTa	SimCSE	36.72	45.50
TiBERT		36.44	39.61
TBERT		42.32	42.93
CINO		51.94	53.27

由表 7 中实验数据可见，1）预训练模型为 RoBERTa 时，SimCSE 模型的相关系数在数据增强前后分别为 36.72% 和 45.50%，提升了 8.78 百分点；预训练模型为 TiBERT 时，SimCSE 模型的相关系数在数据增强前后分别为 36.44% 和 39.61%，提升了 3.17 百分点；预训练模型为 TBERT 时，SimCSE 模型的相关系数在数据增强前后分别为 42.32% 和 42.93%，提升了 0.61 百分点；预训练模型为 CINO 时模型的相关系数在数据增强前后分别为 51.94% 和 53.27%，提升了 1.33 百分点；数据增强后句向量表示模型 SimCSE 的性能均优于数据增强前的性能，说明本文的数据增强方法有效。2）能使 SimCSE 模型性能有明显提升的原

因是：①使用基于形容词知识库的藏文文本数据增强方法使原始数据的数量得到了充分的扩充；②基于形容词知识库的藏文文本数据增强方法所生成的增强句子不仅风格多变，而且表达方式丰富；③通过该数据增强方法生成的扩充句逻辑和语法结构合理。3）数据增强后句向量表示模型 SimCSE 在 4 种预训练模型下性能提升有所不同，其原因是：①预训练模型所使用基元不同，RoBERTa 和 CINO 预训练时的基元为子词，TiBERT 和 TBERT 预训练时的基元为词，子词为基元时不受分词的影响，而词为基元时分词错误将传播到句向量表示，从而影响句向量表示性能；②预训练模型语料大小不同，RoBERTa、TiBERT 和 TBERT 预训练的语料规模分别为 10 GB、3.56 GB 和 6.27 GB，预训练语料规模越大模型性能越佳，从而在 3 种预训练模型下句向量表示性能提升不同。③CINO 在增强前与增强后的 ρ 值均显著高于 RoBERTa、TiBERT 和 TBERT，这一性能优势与其较大的参数量（148~585 MB）相关。通常而言，参数量更大的预训练模型具备更强的特征学习能力，能够捕捉到更丰富、更细致的语言规律与语义关联，因此在句向量表示任务中表现更优。

实验 4　数据增强对文本分类的贡献实验

为了进一步验证数据增强方法在下游任务文本分类中的效果，本实验以文本分类性能较好的卷神经网络（convolutional neural network, CNN）[33]为分类模型，将数据集 TNCC[26] 以 8∶1∶1 的比例分为训练集、验证集和测试集，以精确度（Precision, P）、召回率（Recall, R）和 F1 值（F1）作为评估标准，对比 Repl_Noun 和 Repl_Adj_Obj 两种数据增强方法对分类性能的影响。模型参数设置和 8 实验数据信息见表 8 和表 9。

表 8　模型参数信息

Table 8　Model parameter settings

参数名	值	参数名	值	参数名	值
seq-le	600	filter_sizes	5	rate	10^{-3}
embed-siz	64	num_filters	256	vocab_size	10 000
num_classes	12	dropout	0.8	hidden_dim	128

表 9　文本分类实验数据

Table 9　Experimental results of text classification %

模型	数据增强方法	P	R	F1
CNN	无	64.55	56.90	58.24
	Repl_Noun	67.28	64.86	65.48
	Repl_Adj_Obj	70.52	66.41	67.55

由表 9 中实验数据可见，CNN 模型在基准训练集 TNCC 上的精确率为 64.55%，召回率为 56.9%，F1 值为 58.24%，经 Repl_Noun 方法数据增强后，精确率、召回率和 F1 值分别为 67.28%、64.86% 和 65.48%，依次提升了 2.73、7.96 和 7.20 百分点；经 Repl_Adj_Obj 方法数据增强后，精确率、召回率和 F1 值分别为 70.52%、66.41% 和 67.55%，依次提升了 5.97、9.51 和 9.31 百分点，表明近义词替换法和形容词知识库方法增强数据，藏文文本分类性能都有提高，形容词知识库数据增强方法提升更显著。

4　结束语

本文在分析文本数据增强方法现状的基础上，针对藏文文本数据匮乏的问题，通过分析了藏文文本中形容词的特征，设计了一种基于形容词知识库的藏文文本数据增强模型。模型识别形容词和修饰对象的特征，对形容词和修饰对象进行分类，从而构建形容词知识库，基于形容词知识库和修饰对象近义词表分别替换形容词和形容词修饰对象生成增强文本。本文从数据增强的数量和对下游任务的作用两方面分析了验证基于形容词知识库的藏文文本数据增强方法的有效性，实验数据表明，本文提出的数据增强方法显著扩大了原始数据集的规模，并在藏文句向量表示和文本分类下游任务中性能有明显提升。今后在此研究成果的基础上，进一步探索更多藏文文本数据增强的方法，以不断扩充藏文数据集。

附录

为了便于描述基于形容词知识库的数据增强算法，形容词修饰对象近义词表结构定义如下：

Typedef Struct AdjnearBD //形容词近义词表

{ list adjnear_1;

list adjnear_2;

list adjnear_3;

……

list adjnear_n;}

基于形容词知识库和修饰对象近义词表的文本数据增强算法如下：

算法 1　基于形容词知识库的文本数据增强算法

输入　TiAdjSent //含形容词的藏文句子

输出　TiAdjKBDaSent //数据增强后的藏文句子集

Function Data_augmentation(TiAdjSent)

1) adj_word = getadj(TiAdjSent); //获取形容词

2) adj_object = getobj(TiAdjSent); //获取修饰对象

3) Syntax = getSyntax (TiAdjSent, adj_word); //获取形容词 adj_word 的句式结构

4) if Syntax ∈ SyntaxDB then //若符合 8 种句式结构之一

5) {Repl_Adj (TiAdjSent,adj_word); //替换形容词

6) Repl_ obj (adj_object);｝ //替换修饰对象

7) else Replace_Adj (adj_word); //替换形容词

Function Repl_Adj (TiAdjSent, adj_word)

1) k=located(TiAdj_DB, adj_word);

2) seek(TiAdj_DB,1);

3) m=1;

4) while not eof(TiAdj_DB) //替换形容词

5) {if (TiAdj_DB[k].type.braod=TiAdj_DB[m].type.broad) and

(TiAdj_DB[k].type.sub=TiAdj_DB[m].type.sub) and

(TiAdj_DB[k].syll=TiAdj_DB[m].s yll) and

(TiAdj_DB[k].object.n=TiAdj_DB[m].object.n) and

(TiAdj_DB[k].object.v=TiAdj_DB[m].object.v) and then

6) {sent=replace(TiAdjSent,adj_word,TiAdj_DB[m].*vocab*);

7) append(TiAdjKBDaSent,sent);}

8) m=m+1;}

Function Repl_ object (adj_object)

1) adjnear = located(AdjnearBD, adj_object);

2) sent = replace(TiAdjSent, adj_object, adjnear.words);

3) append(TiAdjKBDaSent,sent).

参考文献：

[1] SHORTEN C, KHOSHGOFTAAR T M, FURHT B. Text data augmentation for deep learning[J]. Journal of big data, 2021, 8(1): 101.

[2] 江荻. 藏语形容词的音节数形态与形态类型[J]. 中国语言学报, 2020(00): 1−27.
JIANG Di. Syllable number morphology and morphological types of Tibetan adjectives[J]. Journal of Chinese linguistics, 2020(00): 1−27.

[3] LITAKE O, YAGNIK N, LABHSETWAR S. IndiText boost: text augmentation for low resource India languages[EB/OL]. (2024−01−23) [2025−03−24]. https://arxiv.org/abs/2401.13085.

[4] 张虎, 张颖, 杨陟卓, 等. 基于数据增强的高考阅读理解自动答题研究[J]. 中文信息学报, 2021, 35(9): 132−140.
ZHANG Hu, ZHANG Ying, YANG Zhizhuo, et al. Data augmentation based automatic answering of reading comprehension in college entrance examination[J]. Journal of Chinese information processing, 2021, 35(9): 132−140.

[5] 葛轶洲, 许翔, 杨锁荣, 等. 序列数据的数据增强方法综述[J]. 计算机科学与探索, 2021, 15(7): 1207−1219.
GE Yizhou, XU Xiang, YANG Suorong, et al. Survey on Sequence Data Augmentation[J]. Journal of frontiers of computer science & technology, 2021, 15(7): 1207−1219.

[6] GHOSH S, TYAGI U, SURI M, et al. ACLM: a selective-denoising based generative data augmentation approach for low-resource complex NER[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto: Association for Computational Linguistics, 2023: 104−125.

[7] YAN Ge, LI Yu, ZHANG Shu, et al. Data augmentation for deep learning of judgment documents[C]//Intelligence Science and Big Data Engineering. Big Data and Machine Learning. Cham: Springer International Publishing, 2019: 232−242.

[8] 王可超, 郭军军, 张亚飞, 等. 基于回译和比例抽取孪生网络筛选的汉越平行语料扩充方法[J]. 计算机工程与科学, 2022, 44(10): 1861−1868.
WANG Kechao, GUO Junjun, ZHANG Yafei, et al. A Chinese-Vietnamese parallel corpus expansion method based on back translation and proportional extraction Siamese network screening[J]. Computer engineering and science, 2022, 44(10): 1861−1868.

[9] ZHANG Jinyi, TIAN Ye, MAO Jiannan, et al. WCC-JC: a web-crawled corpus for Japanese-Chinese neural machine translation[J]. Applied sciences, 2022, 12(12): 6002.

[10] HOANG V C D, KOEHN P, HAFFARI G, et al. Iterative back-translation for neural machine translation[C]//Proceedings of the 2nd Workshop on Neural Machine Translation and Generation. Melbourne: Association for Computational Linguistics, 2018: 18−24.

[11] 祁瑞艳, 李龙杰, 徐世琤, 等. 基于跨度与类别增强的中文新闻命名实体识别[J]. 智能科学与技术学报, 2024, 6(4): 495−508.
QI Ruiyan, LI Longjie, XU Shicheng, et al. Named entity recognition based on span and category enhancement for Chinese news[J]. Chinese journal of intelligent science and technology, 2024, 6(4): 495−508.

[12] ZHOU Chunting, MA Xuezhe, HU Junjie, et al. Handling syntactic divergence in low-resource machine translation[EB/OL]. (2019−08−30)[2025−03−24]. https://arxiv.org/abs/1909.00040.

[13] 廖俊伟. 深度学习大模型时代的自然语言生成技术研究[D]. 成都: 电子科技大学, 2023.
LIAO Junwei. Research on natural language generation techniques in the large language model era of deep learning[D]. Chengdu: University of Electronic Science and

Technology of China, 2023.
[14] ZHANG Xiang, ZHAO Junbo, LECUN Y. Character-level convolutional networks for text classification[J]. Advances in neural information processing systems, 2015: 649−657.
[15] WEI J, ZOU Kai. EDA: easy data augmentation techniques for boosting performance on text classification tasks[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019: 6382−6388.
[16] COULOMBE C. Text data augmentation made simple by leveraging NLP cloud APIs[EB/OL]. (2018−12−05)[2025−03−24]. https://arxiv.org/abs/1812.04718.
[17] FADAEE M, BISAZZA A, MONZ C. Data augmentation for low-resource neural machine translation[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Vancouver: Association for Computational Linguistics, 2017: 567−573.
[18] 张蓉, 刘渊. 适用于方面级情感分析的多级数据增强方法[J]. 数据与计算发展前沿, 2023, 5(5): 140−153.
ZHANG Rong, LIU Yuan. Multi-level data augmentation method for aspect-based sentiment analysis[J]. Frontiers of data & computing, 2023, 5(5): 140−153.
[19] 尤丛丛, 高盛祥, 余正涛, 等. 基于同义词数据增强的汉越神经机器翻译方法[J]. 计算机工程与科学, 2021, 43(8): 1497−1502.
YOU Congcong, GAO Shengxiang, YU Zhengtao, et al. A Chinese-Vietnamese neural machine translation method based on synonym data augmentation[J]. Computer engineering and science, 2021, 43(8): 1497−1502.
[20] 汪超. 基于数据增强技术的藏汉机器翻译方法研究[D]. 拉萨: 西藏大学, 2023.
WANG Chao. A study on Tibetan-Chinese machine translation method based on data enhancement technology[D]. Lasa: Xizang University, 2023.
[21] 色差甲, 班马宝, 才让加, 等. 结合数据增强方法的藏文预训练语言模型[J]. 中文信息学报, 2024, 38(9): 66−72.
SE Chajia, BAN Mabao, CAI Rangjia, et al. Tibetan pre-training language model combined with data enhancement method[J]. Journal of Chinese information processing, 2024, 38(9): 66−72.
[22] 马进武. 藏语语法四种结构明晰[M]. 北京: 民族出版社, 2008.
[23] 吉太加. 现代藏语语法通论[M]. 西宁: 青海民族出版社, 2022.
[24] 马拉毛草. 基于语料库的藏语形容词功能属性研究[D]. 兰州: 西北民族大学, 2013.
MA Lamaocao. Corpus of Tibetan words describe attributes based on function[D]. Lanzhou: Northwest University for Nationalities, 2013.
[25] 周毛太. 藏语形容词的功能分类及其情感研究[D]. 兰州: 西北民族大学, 2020.
ZHOU Maotai. The research on the classification of Tibetan adjectives and it's emotion[D]. Lanzhou: Northwest University for Nationalities, 2020.
[26] QUN Nuo, LI Xing, QIU Xipeng, et al. End-to-end neural text classification for Tibetan[C]//Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Cham: Springer International Publishing, 2017: 472−480.
[27] CER D, DIAB M, AGIRRE E, et al. SemEval-2017 task 1: semantic textual similarity multilingual and Crosslingual focused evaluation[C]//Proceedings of the 11th International Workshop on Semantic Evaluation(SemEval-2017). Vancouver: ACL, 2017: 1−14.
[28] GAO Tianyu, YAO Xingcheng, CHEN Danqi. SimCSE: simple contrastive learning of sentence embeddings[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Punta Cana: Association for Computational Linguistics, 2021: 6894−6910.
[29] SANGJEE D. Sangjeedondrub/tibetan-roberta-basehuggin-gace[EB/OL]. (2024−06−25) [2025−03−24]. https://huggingface.co/sangjeedondrub/Tibetan-roberta-base.
[30] 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和兰州大学开源软件与实时系统教育部工程研究中心. 藏文预训练语言模型 TBERT github [EB/OL]. (2023−10−08)[2025−03−24]. https://github.com/Dslab-NLP/Tibetan-PLM.
[31] LIU Sisi, DENG Junjie, SUN Yuan, et al. TiBERT: Tibetan pre-trained language model[C]//2022 IEEE International Conference on Systems, Man, and Cybernetics. Prague: IEEE, 2022: 2956−2961.
[32] YANG Ziqing, XU Zihang, CUI Yiming, et al. CINO: a Chinese minority pre-trained language model[C]//Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju: COLING, 2022: 3937–3949.
[33] 林荣华. 基于卷积神经网络的句子分类算法[D]. 杭州: 浙江大学, 2015.
LIN Ronghua. Convolutional neural network based sentence classification algorithm. Hangzhou: Zhejiang University, 2015.

作者简介:

仁青吉, 博士研究生, 主要研究方向为藏文信息处理和藏语自然语言处理。E-mail: 1054808891@qq.com。

才智杰, 教授, 博士生导师, 博士, 主要研究方向为藏文信息处理和藏语自然语言处理。发表学术论文 64 篇。E-mail: Czjqhsd@163.com。