·开发研究与工程应用·　　文章编号：1000-3428(2023)06-0292-08　　文献标志码：A　　中图分类号：TP391

基于提示学习的维吾尔语文本分类研究

张博旭，蒲　智，程　曦

（新疆农业大学 计算机与信息工程学院，乌鲁木齐 830000）

摘　要：维吾尔语属于低资源语言和黏着性语言，现有维吾尔语文本分类方法缺少足够的语料来训练维吾尔语预训练模型。因此，维吾尔语无法基于预训练模型提取有效的句向量信息。现有的文本分类方法利用深度学习模型提取词向量，然而，维吾尔语具有特征稀疏且维度偏高的特点，使得其在文本分类上的效果较差。为此，提出基于提示学习的维吾尔语文本分类方法。基于提示学习，采用多语言预训练模型Cino构造不同的模板，利用模型的掩码预测能力对不同的掩码位置进行预测。为避免掩码预测的词汇信息具有多样性，将模板掩盖掉的词向量代替整体的句向量，利用掩码模型的预测能力，以有限大小的向量表示当前句子的语义信息，将下游任务靠近模型的预训练任务，减少在微调阶段两者不同所造成的影响。在爬取维吾尔语网站所构建新闻数据集上进行的文本分类实验结果表明，相比Cino微调预训练模型，融合提示学习的Cino模型的F1值最高可达到92.53%，精准率和召回率分别提升了1.79、1.04个百分点，具有更优的维吾尔语文本分类效果。

关键词：文本分类；维吾尔语；提示学习；预训练模型；深度学习

开放科学（资源服务）标志码（OSID）：

中文引用格式：张博旭，蒲智，程曦.基于提示学习的维吾尔语文本分类研究[J].计算机工程，2023，49(6)：292-299，313.

英文引用格式：ZHANG B X，PU Z，CHENG X.Research on Uyghur text classification based on prompt learning[J].Computer Engineering，2023，49(6)：292-299，313.

Research on Uyghur Text Classification Based on Prompt Learning

ZHANG Boxu，PU Zhi，CHENG Xi

（School of Computer and Information Engineering，Xinjiang Agricultural University，Urumqi 830000，China）

【Abstract】Uyghur，a low-resource and agglutinative language，suffers from insufficient corpus for training pre-existing Uyghur models.This lack hinders the extraction of effective sentence vector information based on pre-training models.Current text classification methods utilize deep learning models to extract word vectors.However，due to the Uyghur language's inherent sparse features and high dimensionality，these methods underperform in text classification tasks.As a response，a Uyghur text classification method based on prompt learning is introduced.Leveraging prompt learning，this paper utilize a multilingual pre-training model，Cino，to construct varied templates and employ the model′s mask prediction ability for predicting different mask positions.To counteract the diversity of lexical information predicted by the mask，the word vector masked by the template is replaced by the entire sentence vector.The predictive ability of the mask model is then used to represent the current sentence's semantic information with a finite size vector.This approach aligns downstream tasks more closely with the model's pre-training tasks，thereby minimizing discrepancies during the fine-tuning stage.Text classification experiments conducted on news datasets，derived from crawled Uyghur websites，demonstrate superior classification performance in Uyghur language texts.Compared to the Cino fine-tuning pre-training model，the fusion prompt learning Cino model yielded the highest F1 value of 92.53%，enhancing accuracy and recall rates by 1.79 and 1.04 percentage points，respectively.

【Key words】text classification；Uyghur language；prompt learning；pre-training model；deep learning

DOI：10.19678/j.issn.1000-3428.0064892

0　概述

当今时代是一个大数据时代，网络上各种各样形式的信息（如文本、语音、图片、视频等）每天都呈井喷式增长，其中，文本信息量级最大。为了有效处理这种信息量较大的文本，研究人员利用高效实用

基金项目：国家自然科学基金（62161048）。

作者简介：张博旭（1998—），男，硕士研究生，主研方向为自然语言处理；蒲　智（通信作者），副教授、博士；程　曦，讲师、博士。

收稿日期：2022-06-02　　**修回日期：**2022-07-29　**E-mail：**320203320@xjau.edu.cn

的自然语言处理（Natural Language Processing，NLP）技术对文本进行分类。

随着BERT[1]（Bidirectional Encoder Representation from Transformers）等预训练模型的提出，NLP领域的各项任务逐渐演变成预训练-微调的范式，通过大量数据训练完成的模型，在很多任务上都达到了非常优秀的分类效果。但是，预训练模型是基于大量语料进行训练，如维吾尔语这种低资源语言，一方面难以获取，另一方面其语言形态复杂，有限开源的大多为多语言预训练模型。在预训练过程中，维吾尔语的语料数量相对较少，造成模型分类效果不是那么理想。在维吾尔语文本分类的早期研究中，大部分都是利用机器学习模型进行分类，如朴素贝叶斯、K近邻和支持向量机等算法。文献[2]提出一种类似于Jaccard相似度的文本和类主题相似度度量方法，实现了相应的维吾尔语分类。文献[3]提出采用TextRank算法和互信息相似度的维吾尔语关键词提取及文本分类方法。随着深度学习的普及，维吾尔语文本分类逐渐形成了提取词向量、传入卷积等模型捕捉语义信息以完成分类的模式。由于维吾尔语的独特性，因此无论是利用机器学习模型还是基于普通深度学习模型的分类方法都难以与预训练模型的分类效果相比。

本文提出基于提示学习的维吾尔语文本分类方法。因维吾尔语语料的匮乏，Cino模型[4]基于多语言预训练模型XLM-R在国内少数民族语言语料上进行2次预训练，使得该模型具有维吾尔语学习的能力。针对维吾尔语在训练过程中语义信息学习不足的情况，本文采用NLP领域的第四范式，即提示学习，融入模板将下游任务重新调整为利用XLM-R的掩蔽语言建模（Masked Language Modeling，MLM）预训练任务的形式，转换成完形填空的任务，并将答案转换成分类的标签，实现维吾尔语文本分类的目的。

1　相关工作

1.1　预训练模型

1.1.1　XLM模型

XLM[5]是Facebook于2019年在Arxiv上提出的模型。尽管当时BERT采用很多语言进行预训练，但是一个模型只能掌握一种语言，在不同语言模型之间无法交互。而XLM模型的提出正好解决了以上问题，因为它是采用多种语言进行训练，所以使模型具备了识别跨语言信息的能力，对于在低资源语言场景下的下游任务，利用XLM来学习其他语料学习到的语言信息。

XLM模型是跨语言模型，共包括了100种语言，而这100种语言共用一个字典，如特殊Token等。字典采用BPE（Byte Pair Encoding）构建。为了使这100种语言分布均匀，XLM采用概率分布方法避免低资源语料在Encoding的过程中被分为单独的Token，具体过程如式（1）和式（2）所示：

$$q_i=\frac{p_i^{\alpha}}{\sum_{j=1}^{N}p_j^{\alpha}} \tag{1}$$

$$p_i=\frac{n_i}{\sum_{k=1}^{N}n_k} \tag{2}$$

其中：α的值为0.5；n代表每个语种的语料数量，以大幅提升数据量少的语种出现的频率。

XLM不同于BERT，它利用了3个预训练任务，因果语言模型（Causal Language Model，CLM）、MLM、TLM（Translation Language Modeling），其中，CLM用于预测给定句子的下一个单词的概率$P(\omega_t|\omega_1,\omega_2,\cdots,\omega_{t-1},\theta)$，而TLM则是一种翻译语言模型，利用平行语料进行有监督训练，实质还是基于MLM任务。本文采用的提示学习方法，也是基于MLM任务。

相比BERT的MLM任务，XLM的MLM任务形式是采用任意个数的句子组成文本，而不是BERT中采取句子对的形式。

1.1.2　RoBERTa模型

RoBERTa[6]为BERT的改进版本，在采用更多的训练数据和更大参数量的同时，还扩大了batch_size。除此之外，RoBERTa还取消了NSP任务，相比BERT的静态MLM任务，RoBERTa使用动态MASK，即不在预处理阶段MASK，每个样本只进行一次随机MASK，在输入模型的过程中动态生成MASK。

文献[7]提出XLM-R，本质结构是XLM+RoBERTa。与XLM不同的是，XLM-R对于α的取值设定为0.3，增加了语种数量和训练数据，采用自监督的方式训练跨语言特征，且不使用Language Embedding。

此外，多语言预训练模型还包括mBART[8]，也就是多语言版本的BART[9]。

1.1.3　Cino模型

虽然XLM-R训练了100种语言，又拓展了训练使用的数据，但是不同语料的语料库大小依旧相差甚远，例如，维吾尔语语料只有0.4 GB，而中文语料却有45 GB左右。在不同语言之间预训练语料数量的差距较大，可能会造成高资源语言对于低资源语言的学习效果产生影响。虽然文献[10]介绍了在部分语种上XLM-R可以获得比BERT更好的分类效果，但是对于低资源语言而言，与近年来人工智能的发展来看，依旧是数据量越大，模型效果越好。

为此，哈尔滨工业大学提出汉语少数民族预训

练语言模型Cino，基于XLM-R进行2次预训练，共使用了6种少数民族语言，并且为了适应少数民族语言，Cino模型还进行了词汇拓展和词汇修剪以减小模型大小。近期，研究人员还推出了base版本和small版本，相比XLM-R，Cino更适合在算力不足的情况下使用，并且在民族语言上可以获得优于XLM-R的效果。

1.2 提示学习

2020年5月GPT-3[11]诞生，其具有1 750亿的参数量，且训练费用达到千万级别的算力。GPT-3可以编剧、写文章、编写代码，其论文称为“Language Models are Few-shot Learners”，重点方向在于少样本学习。因此，GPT-3最大的贡献是利用Prompt做下游任务，打破了传统预训练-微调的形式，不引进多余的参数量，利用模型自身强大的阅读理解能力以及自身学习到的大量语言知识来完成各种各样的任务。相比传统的预训练-微调范式，基于Prompt的训练模式不需要根据任务来定义参数或者引入特定信息，无需太多的标注数据。因此，在需要训练数据较少的情况下，利用提示学习可以把下游任务转换成语言生成任务，这也成为了NLP训练的第四范式。

除了GPT-3强大的语言模型之外，其他模型也能根据自身在预训练阶段学习到的先验知识来进行提示学习。文献[12]提出的小模型利用BERT的预训练任务之一（MLM）进行少样本学习。

XLMRoBERTaMLM实现原理示意图如图1所示。

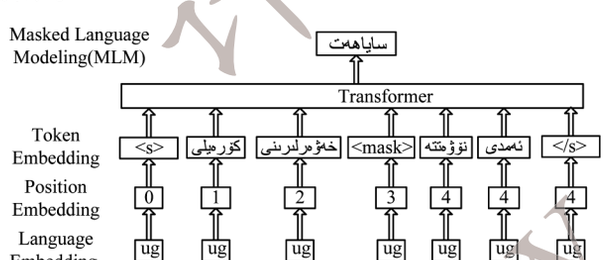

图1 XLMRoBERTaMLM实现原理

Fig.1 Implementation principle of XLMRoBERTaMLM

随机掩码一部分Token，采用“<mask>”标记被遮挡的单词，在经过Embedding层的时候对被遮挡的单词进行向量标记。因为Transformer的映射是多对一的，单个Token的输出向量依赖于整体的向量，所以模型会清楚整句话的内容，从而判断缺少的（即MASK）部分是什么词汇，进而将Token输出的特征向量送入Softmax分类器并输出一个概率分布，根据这个概率分布来判断被遮挡的词汇。但是，在很长一段时间内，MLM都是作为BERT等模型的预训练任务，包括NSP任务。随着研究的不断深入，文献[13]利用MLM完成文本纠错任务，使得MLM任务产生无限的可能性。随着Prompt思想的提出，MLM任务也成了提示学习的一种训练方式。

提示学习又称模板学习，文献[14]介绍了其本质是构建一个模板，即空出一个短语或一个字，与输入的文本相结合，利用掩码模型的预测能力预测空出的短语。提示学习的基本模式如式（3）所示：

$$x'=f_{\text{Prompt}}(x) \tag{3}$$

其中：x为输入的文本；x'为经过f后输出的文本。提示学习的基本模式如下例所示：

新疆是中国的省份，是属于____国的

新疆是中国的省份，是属于[MASK]国的

“是属于____国的”为构建的模板，利用BERT等模型的掩码预测能力可以准确预测出[MASK]位置的词汇为“中”，这种属于单个词汇的MASK。文献[15]提出一个新的框架，利用预定义的跨语言模板进行数据扩充，利用MASK的位置关系捕获跨语言之间的对应关系。由于不同的模型在预训练阶段所使用的编码形式不同，如BERT采用的是字符级BPE编码，而RoBERTa采用byte级别的BPE编码，因此造成Tokenizer后的Token各不相同。不同模型的构造习惯也不尽相同，如BERT两句话的分隔Token为[SEP]，XLM-R的则是</s>，BERT的掩码Token为[MASK]，XLM-R的则为<mask>，BERT每句话的开头Token为[CLS]，XLM-R的则为<s>。因此，在使用不同模型的时候需要确定MASK的形态，以避免模型对于掩码位置识别错误。

当模型利用掩码任务对下例进行预测时：

上星期我去新疆玩的很好，真的_____开心

上星期我去新疆玩的很好，真的[MASK]开心

[MASK]的位置应该是“好”和“不”，但是模型在预测的过程中不单单只会预测出这2个字，如文献[16]介绍，“很开心”可以表达开心的意思，“非常开心”也可以表达开心的意思。因此，当面临诸如此类情感分析问题的时候，一方面可以限定输出，只希望获得“好”和“不”2种情况，另一方面可以做出积极和消极的2种情感标签词映射，让“好”“很”等字代表积极情感，“不”“否”等字代表消极情感，这样就可以通过上下文来预测掩码位置的词汇，将情感分析问题转换成完形填空问题，避免因模型产生多种预测结果而导致最后任务的准确率降低。

模板的选择和位置多种多样，有放在句中也有放在句末的，可以灵活选择。但是模板实质上都是基于人类所学习到的语言知识进行设计的，这种统称为人工设计模板。人工设计出来的模板符合该语言的日常读写习惯，不能太过于生硬，也尽可能不存在语法语序等错误，需要一定的语言掌握能力。文献[17]指出，不同的模板或者标签词对于模型的预测效果影响较大，导致构建的模板不同可能获得的效果各不相同。随着提示学习的不断发展，许多学者开始寻找能够自动构建模板的方式——自动设计

模板,其中包括离散模板。文献[18]介绍了基于挖掘和释义的方式自动构建模板,文献[19]采用挖掘常识知识生成模板。文献[20]摒弃了构造模板的方式,将构建模板转化成连续参数优化的问题,文献[21]则使用梯度引导搜索创建模板。

Prompt的引入可以帮助模型把它所不了解的各种下游任务转换成输出空间有限且熟知的MLM任务,既可以避免参数量过多所造成的微调问题,又可以缩小预训练任务和微调之间的差距,使提示学习更适用于小样本的学习场景。

2 改进的Cino预训练模型

文本分类任务在NLP领域中是一项基础任务,是根据众多已经打好标签的文本对一段全新的文本进行标注的任务。基于文本分类下游任务,目前预训练-微调模型是主流范式,而且对于维吾尔语语种,利用预训练模型生成的词向量或者句向量可以有效降低特征维度,利用预训练的语言能力来获得更优的分类效果。在此之前,文献[22]提出通过稳健词素切分和词干提取等方法,进一步细化词向量并利用长短期记忆(Long Short-Term Memory,LSTM)网络进行特征选择和文本分类。本文基于Cino预训练模型,利用提示学习设计Prompt与每条文本相结合,将输入文本中的每一个字词转换成向量作为模型的输入,通过提取模型输出的词向量,利用全连接层进行特征降维。本文基于MLM预训练任务,将[MASK]掉的位置向量提取出,利用Softmax层输出最大概率的类别,从而实现文本分类。

2.1 模板选择

维吾尔语是一种从右向左读的语言,按照语言习惯,构建模板应该是放在句末。但是,在模型的输入阶段,编码则是按照正常方向读取,如果将模板放在句末,不同的句子生成的Token长度不同,那么无法确定Prompt的位置以及MASK的位置。为验证模板的位置,本文在此之前也采取了Text+Prompt的方式,但是效果却没能达到理想水平,根据以往的维吾尔语实验经验(例如维汉翻译等),无需将维吾尔语按照逆序读取。因此,为了能够清楚地确定是在哪个位置进行掩码,在构建模板的时候,本文采取Text+Prompt的方式,统一采用前置模板。模板选择的详细信息如表1所示。

表1 模板选择的详细信息

Table 1 Details information of template selection

Prompt	翻译	MASK位置
<mask> خەۋەرلىرى	<mask>新闻报导	3
بۈگۈنكى <mask> خەۋەرلىرى	今日<mask>新闻	3
خەۋەر باشتېمىسى <mask>	新闻主题是<mask>	1
ئەمدى دىققەتلەر <mask> خەۋەرلىرىدە بولسۇن	接下来报导一则<mask>新闻	5
ئەمدى نۆۋەتتىكىسى <mask> خەۋەرلىرى	这是一篇<mask>新闻	3
ئەمدى نۆۋەتتە <mask> خەۋەرلىرىنى كۆرەيلى	来看一条<mask>方面的新闻	5
بۇ <mask> ھەققىدىكى خەۋەر	这是一篇关于<mask>的新闻报道	5
ئەمدى دىققەتلەر <mask> خەۋەرلىرىدە بولسۇن	接下来关注一条<mask>新闻	5
ئەمدى <mask> خەۋەرلىرى ئاڭلىتىمىز	插播一条<mask>新闻	6

本文采用9种不同的模板,基于维吾尔语的日常读写习惯来编写,满足新闻报道的文本说明,且长短不相同,MASK的位置也有所不同,包括第一个Token掩码,中间部分Token掩码,尾部Token掩码,基本涵盖了前置模板的各种掩码情况,避免了模板的重复。文献[23]提出因模板细微的不同造成模型分类效果差距过大的问题。本文利用构建模板后的文本分别进行提示学习与预训练模型微调的实验。

2.2 基础模型

本文采用的是Cino预训练模型,由于Cino整体模型架构是基于XLM-R的,因此通过像微调XLM-R或者RoBERTa的形式来微调Cino预训练模型。

2.2.1 Cino预训练模型微调

Cino预训练模型的微调流程如图2所示。

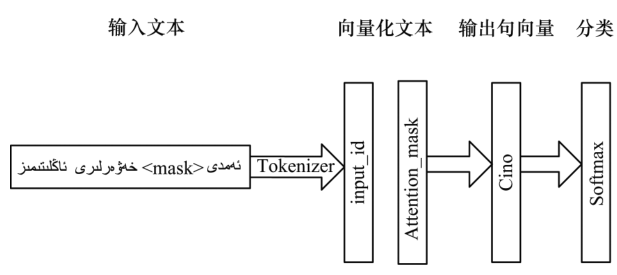

图2 Cino预训练模型的微调流程

Fig.2 Fine tuning procedure of Cino pre-training model

首先利用XLM的分词器将输入的文本进行文本分向量化,一方面生成文本中每个Token对应的词典索引——input_id,另一方面在预训练过程中避免填充符对注意力机制产生影响。因此,将Padding填充的位置设为0,未被Padding填充的位置设为1,生成Attention_mask,这2项就是XLMRoBERTa的

输入形式，将向量化后的文本输入到Cino预训练模型中，提取模型输出的句向量，从而获取整个文本的语义信息，传入Softmax层，将输入的文本信息转换成概率值，提取概率最大的类别，根据标签的映射，即可获知此条语句属于哪个类别。本文构建的维吾尔语文本分类数据集为六分类，预测分类类别的方法如式（4）所示：

$$p(k_i|T,\theta)=\frac{\exp(a_j)}{\sum_{j=1}^{n}\exp(a_j)} \quad (4)$$

其中：i为分类的索引，$i\in\{0,1,2,3,4,5\}$；T为输入的文本；θ为预训练模型所学习到的参数；输出的p为预测最大概率的类别。

2.2.2 对比模型

为了验证提示学习的效果，除了Cino预训练模型之外，本文将Cino预训练模型与卷积神经网络（Convolutional Neural Network，CNN）、循环神经网络（Recurrent Neural Network，RNN）、RCNN、DPCNN 4个模型相结合，构建了多个对比模型Cino_CNN、Cino_RNN、Cino_RCNN、Cino_DPCNN，与提示学习构造的模板进行预训练微调对比。文献[24]利用DPCNN模型对维吾尔语文本进行分类。

TextCNN[25]是一种使用卷积神经网络处理NLP任务的模型，能够更高效地提取文本特征。在预处理阶段，将所有的句子Padding成一个长度输入到模型中，经过Embedding层，加载预训练词向量或者进行随机初始化，再将词向量输入至卷积层中，利用卷积生成特征图，并通过池化层对特征图进行最大池化操作，将拼接最大池化后的特征图传入Softmax层进行归一化处理，得到最终预测值最大的分类类别。TextCNN与预训练模型相结合可以采用Cino输出的词向量来代替预训练的词向量并输入到TextCNN模型中。

除了CNN、RNN、RCNN、DPCNN模型结构不同以外，本文对于文本分类任务的基本操作原理都是一致的，因此只介绍TextCNN模型。

2.3 基于提示学习的Cino预训练模型

本文采用人工构造模板的方式，基于Prompt的特性，模型在输出的过程中，并不是将获取到的句向量输入至Softmax层，也不是将词向量输入到CNN、RNN等模型中再次提取语义信息，而是利用Cino模型自身的掩码预测能力，直接预测Prompt中掩码位置的词汇，根据类别映射或者直接用于分类。但是，这种做法存在以下2个问题：

1）当输入语句为“今天NBA勇士队晋级总决赛，这是一条［MASK］［MASK］新闻”时，采用预训练模型预测MASK位置的词汇，效果好的模型能够精准预测出“体育”2个字。这种预测方式与利用句向量输入至Softmax层的方式有所不同，“运动”“NBA”同样能够说明这条文本是属于体育类别的，因此，无法保证预测输出的文本是类别标签想要的。由于本文的任务属于新闻标签分类，因此无法保证模型能输出多少种和“体育”类别相关的词汇，如果限制模型的输出，就有可能变相降低文本分类的效果。

2）当文本输入模型时文本是被向量化的，需要先经过Tokenize。Cino是基于XLM-R进行2次预训练的，而且维吾尔语属于阿尔泰语系突厥语种，是一种由词干和词缀相结合的复杂语种，形态结构十分丰富，导致维吾尔语在处理过程中原始特征空间维度过大，且特征更加稀疏。在输入模型时，维吾尔语不能采用以每个字为一个Token的分词形式。由于学术界缺少维吾尔语BERT模型，因此不能按照wordpiece方式进行分词。在Tokenize的过程中，Cino是基于RoBERTa进行预训练，利用BPE在5万多byte级别的Token中进行byte到索引的映射，通常频率越高，byte索引越小。

本文在对未精简的维吾尔语文本分类数据集的类别标签进行Tokenize时发现，不同的词汇生成的Token长度是不同的。十分类标签的Token如表2所示。

表2 十分类类别标签的Token

Table 2 Token with ten-group classification category labels

类别标签	Tokenize	Token_len
مائارىپ教育	['▁مائارىپ']	1
ساياھەت旅游	['▁ساياھەت']	1
تۇرمۇش生活	['▁تۇرمۇش']	1
تەنتەربىيە运动	['▁','تەن','تەربىي','ە']	4
قانۇن法律	['▁قانۇن']	1
يېزا ئىگىلىكى农业	['▁يېزا','▁ئىگىلىك','ى']	3
كۆڭۈل ئېچىش娱乐	['▁كۆڭۈل','▁ئېچى','ش']	3
ئىقتىساد经济	['▁ئىقتىساد']	1
ئاپتوموبىل汽车	['▁ئاپتوموبىل']	1
پەن-تېخنىكا科技	['▁پەن','-','تېخنىك','ا']	4

从表2可以看出，不同的维吾尔语标签的Token长度不相同，中文就是2个字长度的词汇，然而，采用维吾尔语写出后编码则有可能是4个字长度，这就是byte级别BPE编码所产生的效果。当模型在掩码预测过程中，由于Prompt所MASK的长度是一定的，因此预测出的Token也是一定的，原本4个Token的标签只能预测一个MASK，不仅输出的语义信息不准确，而且多个Token也无法设定一个合适的标签映射，对结果造成很大的影响。因此，有效地动态设计MASK的长度或限定Token长度的标签成为解决此类问题的方法。

基于以上问题，本文对提示学习进行改进，将微调与提示学习相结合，将最初构建的十分类维吾尔语数据集减少到六分类，基于维吾尔语的日常读写

习惯，舍弃经过Tokenizer后长度大于1的Token所对应的标签数据，以确保六分类的每个类别标签生成的Token长度都为1，在设计的Prompt中统一只掩码一个位置的Token，利用掩码模型的预测能力降低模型预测输出的空间。在计算损失函数时，交叉熵损失函数如式(5)所示：

$$L(\boldsymbol{x},\text{class})=-\log_a\left(\frac{\exp(\boldsymbol{x}[\text{class}])}{\sum_i\exp(\boldsymbol{x}[i])}\right)=$$
$$-\boldsymbol{x}[\text{class}]+\log_a\left(\sum_i\exp(\boldsymbol{x}[i]\right) \quad (5)$$

其中：$\boldsymbol{x}$为网络的最后一层输出，经过全连接层处理的句向量。本文将其修改为掩码模型输出MASK位置的词向量，设计的模板是MASK一个位置的Token，则输出的词向量也是一个词的词向量。同样，经过全连接层的处理，将输出MASK位置的Token转化为输出各个标签最大概率的分类问题，采用掩码预测的词向量代替整体的句向量语义信息。因为一句话的语义信息在文本分类任务的场景中等同于标签的语义信息，所以Prompt所MASK的部分正是标签所代表的词向量，蕴含着标签的语义信息，因此，将这部分的词向量看作整体句子向量信息的集合。在预测过程中，利用torch.argmax()函数单独将MASK部分的词向量提取出放入Softmax层，dim值设置为1，取行的最大值的index，以输出最大概率的标签，从而有效避免了因模型输出多种Token而分类困难的问题。

除此之外，Prompt最初的使用情况就是零样本学习，为了验证在维吾尔语zero-shot实验上Prompt的表现，本文还额外设计了zero-shot实验，不经过预训练-微调，直接利用Prompt来微调模型进行文本分类，测试模型利用提示学习在无监督环境下的学习能力。

3　实验

3.1　实验环境

实验平台为托管式Jupyter笔记本Colab以及Linux操作系统，硬件为AMD EPYC 7302处理器，8核CPU，64 GB内存，NVIDIA GeForce RTX 3090显卡，24 GB显存。模型采用Python语言实现，Python版本为3.8，使用的深度学习代码库包括PyTorch1.11.0、sklearn1.0.2、transformers4.19.2、numpy1.21.6、sentencepiece0.1.96。编码实现工作通过Colab完成。

3.2　数据集

由于目前缺少开源的维吾尔语文本分类数据集，因此本文采用爬虫技术，通过对天山网、中亚之声、中国维吾尔语广播网、中国喀什网、昆仑网|新疆党建网、阿克苏新闻网、人民网维吾尔文、Nur网8个网站进行新闻标题的爬取，共32 327条语料，涉及十分类。为了避免生成的Token长度不统一问题，本文对语料进行裁剪，最终生成17 910条语料，共六分类，包括教育(1 408条)、旅游(2 344条)、生活(5 907条)、法律(251条)、经济(4 000条)、汽车(4 000条)，并进行一系列的预处理操作，如去除停用词、去除乱码等操作。本文根据训练集∶验证集∶测试集为3∶1∶1的比例，对数据集中每个类别都依照此比例进行划分。

3.3　超参数设置

Cino预训练模型的具体超参数设置如表3所示。

表3　Cino预训练模型超参数设置

Table 3　Hyperparameter settings of Cino pre-training model

超参数	值或函数
学习率	5×10^{-6}
Epoch	2
gradient_acc	8
batch_size	4
Max_len	512
Weight_decay	1×10^{-4}
warmup_rate	0.1
Dropout(CNN、RNN、RCNN、DPCNN)	0.1

由于实验室算力有限，且基于掩码模型训练的参数量过大，维吾尔语文本经过Tokenizer的长度普遍偏长，因此将batch_size设置为4，在encoding阶段将Padding设置为Max_len(512)。但是，batch_size设置过小可能会对模型的分类效果产生一定的影响，因此，可以使用梯度累加的方法解决batch_size的设置问题，如式(6)所示：

loss/=self.config.gradient_acc　　(6)

由于loss每次都除以gradient_acc，因此当累加到一定步数时optimizer更新的值实际上是这些累加值的均值，变相地提升了batch_size，使更新后的loss更加平稳。除此之外，本文利用训练热身(warmup)动态调整学习率，并且为了避免过拟合现象，使用权值衰减(Weight_decay)，调节模型复杂度对损失函数的影响。经过多次实验对比分析，Cino预训练模型在训练3个Epoch的时候就已经达到拟合状态，损失值不再上升，故将Epoch设置为2。

3.4　评价指标

本文采用sklearn库中的classification_report库输出的评价指标精确率(P)、召回率(R)、F1值，以及准确率(A)作为此次文本分类实验的评价指标。六分类的混淆矩阵如表4所示，其中，TP(True Positive)为正面预测正确，TN(True Negative)为反面预测正确，FP(False Positive)为正面预测错误，FN(False Negative)为反面预测错误。

表4 一个类别的混淆矩阵

Table 4 Confusion matrix of one category

类别	教育	旅游	生活	法律	经济	汽车
教育	—	—	—	—	—	FP
旅游	—	—	—	—	—	FP
生活	—	—	—	—	—	FP
法律	—	—	—	—	—	FP
经济	—	—	—	—	—	FP
汽车	FN	FN	FN	FN	FN	TP

在混淆矩阵中，正确的分类样本总存在于从左上到右下的对角线之中。准确率定义为分类正确的样本数量，也就是对角线上的样本数量与总样本数量的比值。由于本文描述的是多分类问题，准确率所计算的是全局的样本预测情况，因此采用二分类的准确率计算公式来代替多分类问题，如式(7)所示：

$$A=\frac{T_{\mathrm{TP}}+F_{\mathrm{FN}}}{T_{\mathrm{TP}}+T_{\mathrm{TN}}+F_{\mathrm{FP}}+F_{\mathrm{FN}}} \tag{7}$$

每次计算都单独计算各个类别的精确率和召回率，表4中以“汽车”类别为例，精确率和召回率如式(8)和式(9)所示：

$$P=\frac{T_{\mathrm{TP}}}{T_{\mathrm{TP}}+F_{\mathrm{FP}}} \tag{8}$$

$$R=\frac{T_{\mathrm{TP}}}{T_{\mathrm{TP}}+F_{\mathrm{FN}}} \tag{9}$$

F1值是精确率和召回率的调和平均数，如式(10)所示：

$$\mathrm{F1}=\frac{2\times P\times R}{P+R} \tag{10}$$

在面临多分类问题的情况下，还存在3种计算各项指标平均值的方法：1)宏平均对每个类别的精确率、召回率、F1值做加和求平均值；2)微平均(micro avgrage)不区分类别，计算整体的精确率、召回率、F1值；3)加权平均(Weight avgrage)则是对宏平均的进一步改进，考虑每个类别数量在总样本数量中的占比。结合此次实验自建数据集的特性和实验结果，本文采用宏平均方式计算各种评价指标的平均值，以下实验结果中的各项指标数据均是宏平均后的结果。

3.5 实验结果

本文首先通过Cino预训练模型进行微调，在Cino预训练模型的基础上增加CNN、RNN、RCNN、DPCNN 4种模型结构进行实验，其次利用提示学习构建9种不同的模板再次进行微调实验，最后利用这些模板进行零样本学习实验。不同模型的对比实验结果如表5所示，加粗表示最优数据。从表5可以看出：基础Cino预训练模型的准确率最高，达到了94.55%，而加入了CNN结构后，Cino_CNN模型在精确率和F1值上都有了不同程度的提升，加入RCNN则进一步提升了召回率。因此，卷积结构相对于句向量可以进一步提取语义信息。

表5 不同模型的评价指标对比

Table 5 Evaluation indicators comparison among different models %

模型	精确率	召回率	F1值	准确率
Cino	93.04	90.83	91.71	**94.55**
Cino_CNN	**93.87**	90.85	**92.14**	94.41
Cino_DPCNN	93.57	89.81	91.33	93.76
Cino_RCNN	92.67	**91.43**	91.76	94.38
Cino_RNN	93.40	90.77	91.76	93.96

在Cino预训练模型的基础上融合提示学习的实验结果如表6所示(由于模板长度较长，因此根据表1的Prompt顺序将各个模板缩写成Prompt 1~Prompt 9)。

表6 在Cino预训练模型的基础上融合提示学习的实验结果

Table 6 Experimental results of integrating prompt learning on the basis of Cino pre-training model %

模板	精确率	召回率	F1值	准确率
Prompt 1	93.53	**91.87**	**92.53**	**94.80**
Prompt 2	93.37	90.92	91.84	93.96
Prompt 3	93.61	90.63	91.87	94.18
Prompt 4	**94.83**	90.69	92.51	94.49
Prompt 5	93.72	90.90	92.02	93.87
Prompt 6	94.09	90.38	91.99	94.32
Prompt 7	92.92	89.41	90.74	93.37
Prompt 8	93.83	91.27	92.31	94.43
Prompt 9	94.43	90.91	92.44	94.35

从表6可以看出：Prompt 1在召回率、F1值、准确率方面效果最好，Prompt 4的精确率最高，除Prompt 7以外，其余所有的模板都获得了比直接微调Cino预训练模型更高的F1值。但是，本文设计的不同模板具有不同的分类效果，在各模板中F1值的最大值和最小值相差1.79个百分点，因此，人工设计模板存在一定的困难和不确定性。融合表5和表6中表现最优的模型和模板进行对比，对比结果如表7所示。

表7 最优实验结果对比

Table 7 Comparison of optimal experimental results %

模型或模板	精确率	召回率	F1值	准确率
Cino	93.04	90.83	91.71	94.55
Cino_CNN	93.87	90.85	92.14	94.41
Cino_RCNN	92.67	91.43	91.76	94.38
Prompt 1	93.53	**91.87**	**92.53**	**94.80**
Prompt 4	**94.83**	90.69	92.51	94.49

从表7可以看出：在预训练模型的基础上利用提示学习构建模板，无论是在精确率、召回率、F1值还是准确率上都有一定程度的提升，验证了利用模

板进行微调的可行性，也证明了MLM任务具有的强大能力。本文还利用提示学习简单做了一些零样本实验，测试Prompt在无监督环境下的表现情况，具体实验结果如表8所示。

表8　零样本实验结果

Table 8　Experimental result of zero-sample　%

模型或模板	精确率	召回率	F1值	准确率
Cino	0.22	16.66	0.44	1.34
Prompt 1	20.42	16.69	6.18	22.41
Prompt 2	0.46	11.18	0.43	1.03
Prompt 3	2.18	16.66	3.87	13.13
Prompt 4	3.73	16.66	6.09	22.35
Prompt 5	1.75	16.51	2.51	7.85
Prompt 6	**34.27**	**17.00**	6.69	10.31
Prompt 7	5.49	16.66	**8.26**	**32.95**
Prompt 8	0.22	16.66	0.44	1.34
Prompt 9	3.69	16.99	4.46	12.68

从表8可以看出：在微调Cino预训练模型的基础上加入提示学习模板的大部分实验结果都得到有效提升，相比Cino，加入Prompt模板构建模型的精确率最高可提升34个百分点，虽然F1值最高只有8.26%，准确率最高只有32.95%，但是也侧面验证了加入模板之后，模型对于从未见过的任务和数据，利用预训练阶段掩码预测的能力对模型从未见过的下游任务进行处理。文献[26]利用ZeroPrompt拓展1 000个预训练任务来进行零样本泛化，并大幅提升了文本分类效率。

4　结束语

本文提出基于提示学习的维吾尔语文本分类方法。通过将提示学习与微调预训练模型相结合，利用预训练模型的掩码预测能力，使下游任务和预训练阶段的任务相靠近，缓解维吾尔语这种黏着性语言造成的特征稀疏问题。通过预测文本中模板部分被掩码掉的词向量，并将其单独提取出来输入到Softmax层，以低维度的向量代替高维度的句向量，从而精简整体语义信息并进行文本分类。在自建维吾尔语文本分类数据集上的实验结果验证提示学习的有效性，实验结果表明，本文方法能够有效提升维吾尔语文本分类的准确性。考虑到人工设计模板具有一定的复杂性和不确定性，后续将继续基于维吾尔语进行自动设计模板的实验，利用参数不断更新优化代替基于自然语言的人工模板，并将研究重点转移到零样本实验中，进一步提升维吾尔语文本分类在零样本实验上的效果。

参考文献

[1] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2022-04-28]. https://arxiv.org/pdf/1810.04805.pdf.

[2] 吐尔地·托合提，维尼拉·木沙江，艾斯卡尔·艾木都拉. 基于语义串抽取及主题相似度度量的维吾尔文文本分类[J]. 中文信息学报，2017，31(4)：100-107.
Turdi Tohti, Winira Musajan, Askar Hamdulla. Semantic string-based topic similarity measuring approach for uyghur text classification[J]. Journal of Chinese Information Processing, 2017, 31(4): 100-107. (in Chinese)

[3] 阿力甫·阿不都克里木，李晓. 基于TextRank算法和互信息相似度的维吾尔文关键词提取及文本分类[J]. 计算机科学，2016，43(12)：36-40.
Ghalip Abdukerim, LI X. Uyghur keyword extraction and text classification based on TextRank algorithm and mutual information similarity[J]. Computer Science, 2016, 43(12): 36-40. (in Chinese)

[4] YANG Z Q, XU Z H, CUI Y M, et al. Cino: a Chinese minority pre-trained language model[EB/OL]. [2022-04-28]. https://arxiv.org/abs/2202.13558.

[5] LAMPLE G, CONNEAU A. Cross-lingual language model pretraining[EB/OL]. [2022-04-28]. https://arxiv.org/pdf/1901.07291.pdf.

[6] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach[EB/OL]. [2022-04-28]. https://arxiv.org/abs/1907.11692.

[7] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale[EB/OL]. [2022-04-28]. https://arxiv.org/abs/1911.02116v2.

[8] LIU Y H, GU J T, GOYAL N, et al. Multilingual denoising pre-training for neural machine translation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 726-742.

[9] LEWIS M, LIU Y H, GOYAL N, et al. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[EB/OL]. [2022-04-28]. https://arxiv.org/abs/1910.13461.

[10] WU S J, DREDZE M. Are all languages created equal in multilingual BERT?[EB/OL]. [2022-04-28]. https://arxiv.org/abs/2005.09093.

[11] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in Neural Information Processing Systems, 2020, 33: 1877-1901.

[12] SCHICK T, SCHÜTZE H. It's not just size that matters: small language models are also few-shot learners[EB/OL]. [2022-04-28]. https://arxiv.org/abs/2009.07118? utm_medium=email&_hsenc=p2ANqtz-_QwAkpWYd5cbmMTX5gb9_GYEBsWkI_vi0WyIti1i3vzXI7Qw0zTGiLe6VfcuW-v15PRAlZ.

[13] ZHANG S H, HUANG H R, LIU J C, et al. Spelling error correction with soft-masked BERT[EB/OL]. [2022-04-28]. https://arxiv.org/abs/2005.07421v1.

[14] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[EB/OL]. [2022-04-28]. https://arxiv.org/abs/2107.13586.

[15] QI K, WAN H, DU J, et al. Enhancing cross-lingual natural language inference by prompt-learning from cross-lingual templates[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2022: 1910-1923.

（下转第313页）

(上接第299页)

[16] GRIVAS A, BOGOYCHEV N, LOPEZ A. Low-rank softmax can have unargmaxable classes in theory but rarely in practice[EB/OL]. [2022-04-28]. https://arxiv. org/abs/2203. 06462v2.

[17] GAO T Y, FISCH A, CHEN D. Making pre-trained language models better few-shot learners [EB/OL]. [2022-04-28]. https://arxiv. org/abs/2012. 15723.

[18] JIANG Z B, XU F F, ARAKI J, et al. How can we know what language models know? [J]. Transactions of the Association for Computational Linguistics, 2020, 8: 423-438.

[19] DAVISON J, FELDMAN J, RUSH A. Commonsense knowledge mining from pretrained models[C]//Proceedings of Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2019: 1173-1178.

[20] LIU X, JI K X, FU Y C, et al. P-Tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks[EB/OL]. [2022-04-28]. https://arxiv. org/abs/2110. 07602v2.

[21] SHIN T, RAZEGHI Y, LOGAN IV R L, et al. Autoprompt: eliciting knowledge from language models with automatically generated prompts[EB/OL]. [2022-04-28]. https://arxiv. org/abs/2010. 15980v1.

[22] 沙尔旦尔·帕尔哈提，米吉提·阿不里米提，艾斯卡尔·艾木都拉. 基于稳健词素序列和LSTM的维吾尔语短文本分类[J]. 中文信息学报，2020，34(1)：63-70.
Sardar Parhat, Mijit Ablimit, Askar Hamdulla. Uyghur short text classification based on robust morpheme sequence and LSTM [J]. Journal of Chinese Information Processing, 2020, 34(1): 63-70. (in Chinese)

[23] LIU X, ZHENG Y N, DU Z X, et al. GPT understands, too [EB/OL]. [2022-04-28]. https://arxiv. org/abs/2103. 10385v1.

[24] 加米拉·吾守尔，吴迪，王路路，等. 基于多卷积核DPCNN的维吾尔语文本分类联合模型[J]. 中文信息学报，2021，35(7)：63-71.
Jiamila Wushouer, WU D, WANG L L, et al. Uyghur text categorization joint model based on multi-convolution kernel DPCNN [J]. Journal of Chinese Information Processing, 2021, 35(7): 63-71. (in Chinese)

[25] YOON K. Convolutional neural networks for sentence classification[EB/OL]. [2022-04-28]. http://de. arxiv. org/pdf/1408. 5882.

[26] XU H W, CHEN Y J, DU Y L, et al. Zeroprompt: scaling prompt-based pretraining to 1,000 tasks improves zero-shot generalization[EB/OL]. [2022-04-28]. https://arxiv. org/abs/2201. 06910.

编辑　薛晋栋