TiKEM: 基于知识增强的藏文预训练语言模型

邓俊杰[1,3] 陈龙[1,3] 张廷[1,2,3] 孙媛[1,3,4,*] 赵小兵[1,3,*]
[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，赵小兵
tracy.yuan.sun@gmail.com

摘要

预训练语言模型在中英文领域有着优异的表现，而低资源语言数据获取难度大，预训练语言模型在低资源语言如藏文上的研究刚取得初步进展。现有的藏文预训练语言模型，使用大规模无结构的文本语料库进行自监督学习，缺少外部知识指导，知识记忆能力和知识推理能力受限。为了解决以上问题，本文构建含有50万个三元组知识的藏文知识增强预训练数据集，联合结构化的知识表示和无结构化的文本表示，训练基于知识增强的藏文预训练语言模型TiKEM，以提高模型的知识记忆和推理能力。最后，本文在文本分类、实体关系分类和机器阅读理解三个下游任务中验证了模型的有效性。

关键词： 藏文；知识增强；预训练语言模型；文本分类；实体关系分类；机器阅读理解

TiKEM: Knowledge Enhanced Tibetan Pre-trained Language Model

Junjie Deng [1,3] Long Chen[1,3] Ting Zhang[1,2,3] Yuan Sun [1,3,4,*] Xiaobing Zhao[1,3,*]
[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 and Xiaobing Zhao
tracy.yuan.sun@gmail.com

Abstract

The pre-trained language model has excellent performance in the Chinese and English fields. But the research of pre-trained language models in low-resource languages such as Tibetan have just made initial progress. The main reason is that it's difficult to obtain low-resource language data. The existing Tibetan pre-trained language model uses a large-scale unstructured text corpus for self-supervised learning, which lacking external knowledge guidance. So their knowledge memory and knowledge reasoning abilities are limited. To solve the above problems, this paper build a Tibetan knowledge enhancement pre-trained dataset containing 500,000 triples of knowledge. Then, this paper combine structured knowledge representation and unstructured text representation to train the knowledge enhanced Tibetan pre-trained language model TiKEM. This method can improve the knowledge memory and reasoning abilities of the model. Finally, this paper verifies the effectiveness of the model in Text Classification, Tibetan relationship classification and Tibetan machine reading comprehension.

Keywords: Tibetan , Knowledge Enhancement , Pre-trained Language Model , Text Classification , Entity Relationship Classification , Machine Reading Comprehension

第二十二届中国计算语言学大会论文集，第135页-第144页，哈尔滨，中国，2023年8月3日至5日。

1 引言

预训练语言模型可以从大规模无标签的数据中学习丰富的上下文表征，通过迁移学习在多个下游任务上取得了优秀的性能，这对于低资源语言上的自然语言处理研究存在重要意义。目前，预训练语言模型在英文等语言上取得了很好的发展，一系列预训练语言模型相继出现，如GPT(Radford et al., 2018)、BERT(Kenton and Toutanova, 2019)和RoBERTa(Liu et al., 2019)等。为了进一步优化预训练语言模型，研究人员通过增加模型参数量，提高预训练语言模型在下游任务中的性能，如GPT-3(Brown et al., 2020)、T5(Raffel et al., 2020)、盘古α(Zeng et al., 2021)等。模型参数量的增加，虽然显著提升了模型性能，但是在知识获取能力方面依然存在不足。对于该问题，研究人员将知识图谱中的事实知识融合到预训练语言模型中，如ERNIE(Zhang et al., 2019)、KEPLER(Wang et al., 2021)、ERNIE3.0(Sun et al., 2021a)等，显著提高了模型的认知能力。

以上研究均是在英文等资源比较丰富的语言上的研究，为了解决在低资源语言上数据稀疏等问题，人们提出多语言预训练模型mBERT(Pires et al., 2019)、XLM-R(Conneau et al., 2020)等，但上述多语言预训练模型在藏文上的效果并不理想，例如mBERT在藏文分类数据集TNCC(Qun et al., 2017)上的F1为5.5%，XLM-R-base的F1为21.1%(Yang et al., 2022)。为此，Liu等人(Liu et al., 2022)提出藏文预训练语言模型TiBERT，Yang等人(Yang et al., 2022)提出了少数民族多语言预训练模型CINO，Deng等人(Deng et al., 2023)提出了少数民族多语言预训练语言模型MiLMo。以上模型推动了藏文自然语言处理的研究，但目前的藏文预训练语言模型都是使用大规模无标注数据进行自监督学习，缺少外部知识指导，知识记忆能力和知识推理能力存在不足。

针对上述存在的不足，本文提出基于知识增强的藏文预训练语言模型TiKEM，联合结构化知识和无结构化文本，并在下游任务中评估模型性能，主要贡献如下：

（1）为了显示表示知识，本文构建了一个含有50万个三元组的藏文知识库，并将其与藏文语料库结合，构建藏文知识增强预训练数据集；

（2）为了提高藏文预训练语言模型的知识记忆和知识推理能力，本文提出基于知识增强的藏文预训练语言模型TiKEM，统一建模结构化知识表示和无结构文本表示，对知识进行掩码。同时，为了增强模型的句子表达能力，本文将下一个句子预测任务扩展为句子重排序任务和句子间的距离关系任务；

（3）为了评估模型的性能，本文在文本分类、实体关系分类、阅读理解三个下游任务中进行了对比实验，实验结果表明本文提出的藏文预训练语言模型的性能有着显著的提升。

2 相关工作

预训练语言模型已经在自然语言处理的多项下游任务中取得了优秀的性能，包括OpenAI GPT(Radford et al., 2018)、BERT(Kenton and Toutanova, 2019)、XLNet(Yang et al., 2019)等，可以有效获取句法和语义信息来进行文本表示。尽管预训练语言模型的上下文表示已经包含了句法、语义等知识，但挖掘上下文表示所蕴含的知识的研究较少，它对于文本理解非常重要。Zhou等人(Zhou et al., 2020)在不同具有挑战性的测试中检验GPT、BERT、XLNet和RoBERTa的常识获取能力，发现模型在需要更多深入推理的任务上表现不佳，这也表明常识获取依然是一个巨大挑战。

知识图谱存储着丰富的知识，利用知识图谱让模型显式学习人类对世界的认知，是融合知识的预训练模型采用的重要方法(王海峰et al., 2022)。该方法可以提高预训练模型的知识获取和知识推理等能力。ERNIE(Zhang et al., 2019)首先对文本中提到的命名实体进行识别提取，然后将实体与知识图中对应的实体对齐，利用文本语义作为知识图的实体嵌入，再使用TransE方法学习图的结构。然后利用掩码机制，将知识图中的实体遮蔽，使模型聚合上下文和知识图共同预测遮掩的令牌和实体，使得预训练模型不仅可以将图谱三元组中的事实知识更好地融合到模型中，而且还可以通过丰富的实体描述，有效地学习实体和关系的知识表示。其提供了整合异构数据的一种示范方法。Sun等人(Sun et al., 2020)认为，通过多种不同的知识表示学习获得的实体嵌入，并在预训练阶段进行融合的方法，不能够充分学习到相应知识，并且当知识图谱发生变化时需要重新训练实体嵌入表示模型。因此在CoLAKE模型中提出词-知识图的

概念，将文本序列看作是全连接的词图，以构成一个同时包含词语、实体和关系的词语-知识图。CoLAKE利用遮蔽注意力来控制信息流，将掩码策略分为词节点掩码、实体节点掩码、关系节点掩码，从而能够同时融合训练语料中的语言知识和图谱中的知识。然而，CoLAKE更加关注于实体在知识图谱中的建模，却忽略了实体在训练语料中的表述，在一定程度上削弱了语言模型的泛化能力。为此，ERNIE3.0(Sun et al., 2021a)提出知识图谱与文本平行预训练的方法，使用文本表述知识。其将知识图谱中的三元组与对齐文本统一编码，作为预训练语言模型的输入，同时利用掩码策略，掩盖三元组中的关系和文本中的实体，促使模型融合三元组知识和文本信息。

在藏语方面，哈工大讯飞联合实验室提出了包含藏语、蒙语（回鹘体）、维吾尔语、哈萨克语（阿拉伯体）、朝鲜语、壮语、粤语七种少数民族语言的多语言预训练模型CINO(Yang et al., 2022)，该模型基于多语言预训练模型XLM-R(Conneau et al., 2020)开发，在多种少数民族语料上进行了二次预训练，该模型在藏文分类数据集TNCC(Qun et al., 2017)上相比其它基线模型获得了显著的性能提升。Liu等人(Liu et al., 2022)提出了藏文预训练语言模型TiBERT，其构建了覆盖语料库99.95%的词汇表。该模型在文本分类和问题生成任务中取得较好效果。Deng等人(Deng et al., 2023)提出了包含蒙古语、藏语、维吾尔语、哈萨克语和韩语五种少数民族语言的多语言预训练模型MiLMo，该模型在构建的多语言文本分类数据集上证明了模型的有效性，推动了少数民族语言信息化的建设。安波等人(安波and 龙从军, 2022)提出了藏文预训练语言模型BERT-base-Tibetan，并将该模型应用于藏文文本分类，实验表明预训练语言模型能显著提升藏文文本分类的性能。目前的藏文预训练语言模型取得了不错的进展，但大都是使用大规模无标注数据进行自监督学习，缺少外部知识指导，知识记忆能力和知识推理能力存在不足。因此如何使用知识库来增强藏文预训练模型的表示能力是藏文预训练模型研究和应用的难点之一。

3 TiKEM模型

本文构建的藏文知识增强预训练语言模型总体结构如图1所示。首先将知识库中的三元组与语料库中的文本进行拼接作为训练数据。本文使用[SEP]分隔三元组与文本并将其作为预训练语言模型的输入，不使用额外的模型对知识库中的三元组进行知识表示，从而统一建模结构化知识表示和无结构化文本表示。然后分别根据实体掩码预测任务、知识掩码预测任务、句子重排序任务和句子间的距离关系任务的要求，对输入文本进行处理。如知识掩码预测任务中，随机掩码三元组中的关系或文本中的实体。TiKEM与CharBERT(Ma et al., 2020)结构类似，融合子词与字符表示，使用多层双向Transformer作为通用知识文本表示。在此基础上，使用多任务学习框架，分别对四个任务进行建模，最后以四个任务的加权损失函数值作为模型总体损失值。

3.1 预训练任务

基于掩码语言模型和下一个句子预测任务，本文使用实体掩码预测、知识掩码预测、句子重排序和句子间的距离关系共四个任务作为预训练任务，具体如下。

3.1.1 实体掩码预测

BERT的掩码预测任务是预训练语言模型最重要的预训练任务，它可以帮助模型更好地理解语言中的上下文和语义信息。在BERT的掩码预测任务中，模型需要预测输入文本中被随机遮掩的子词，但是随机掩码子词会影响全词表示，不利于模型对实体、单词、短语的理解。因此本文随机选取输入文本中15%的令牌进行掩码，其中实体占80%，即在文本中代表具体事物、人物、地点等的词语，而剩余20%的令牌则是非实体的单词或短语。在这个过程中，为了提高模型鲁棒性，80%的令牌使用特殊标记[MASK]进行替换，10%的令牌随机替换为其它令牌，剩余10%的令牌不进行处理。掩码语言模型的损失值$loss_{mlm}$即为其损失值，如公式（1）所示。

$$loss_{mlm} = -\sum_{i=1}^{n} p_m \log p'_m \tag{1}$$

其中，p'_m为预测的token。p_m为真实token。

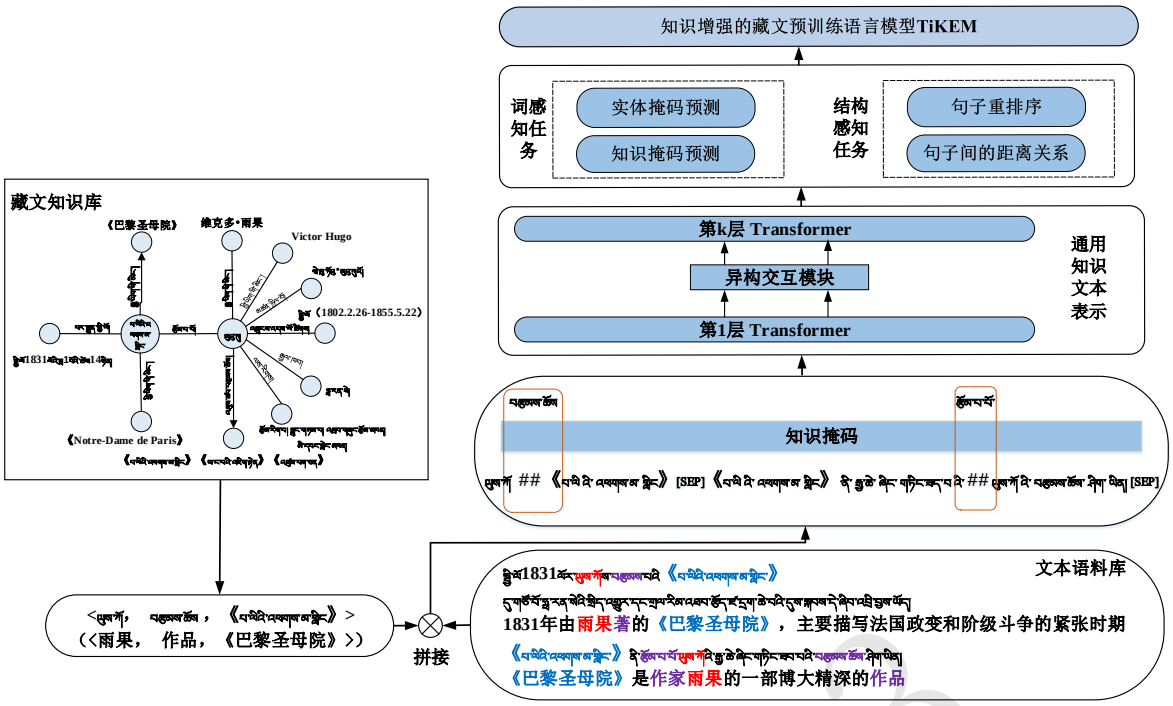

图 1: 基于知识增强的藏文预训练语言模型TiKEM

3.1.2 知识掩码预测

知识掩码预测任务是指将知识库中的三元组与给定文本作为预训练模型的序列输入，以特殊标记[SEP]隔开，并随机掩码三元组中的关系或文本中对应的实体，让模型结合三元组知识与文本知识，预测三元组的关系或文本中的实体,与实体掩码预测相同，掩码语言模型的损失值$loss_{mlm}$同样为其损失值。

在传统的语言模型中，模型需要学习词汇之间的关系和语言结构，而在知识掩码预测任务中，模型需要学习知识库中的三元组关系，并将其与文本融合，从而增强模型对知识的理解和应用能力。因此本文将藏文知识库中的三元组与藏文语料库中的文本使用“[SEP]”进行拼接，作为序列输入。其中三元组以主体-关系-客体的形式呈现，如图1所示，（ཡུས་ཀོ （雨果），བརྩམས་ཚོས （创作），པ་ལིའི་འཐབས་མ་སྒྲིང （巴黎圣母院））。本文在序列中随机掩码三元组中的关系或文本中对应的实体，例如在图1中，三元组中的“ བརྩམས་ཚོས （创作）”以及语料中的“ རྩོམ་པ་པོ་ （作家）”被遮掩，这促使模型结合三元组知识与文本蕴含知识，预测三元组的关系或文本中的实体，从而学习三元组中主客体之间的关系和知识。其本质类似于实体关系抽取中的远程监督算法(Mintz et al., 2009)。远程监督算法假设如果两个实体参与了一个关系，那么任何包含这两个实体的句子都可以表达这种关系。

与传统语言模型中的掩码预测任务相比，知识掩码预测任务使模型能够显示学习人类对世界的认知，从而更好地应对自然语言处理中的任务和问题。

3.1.3 句子重排序

句子重排序任务是将给定的输入序列分成多段文本，并随机打乱，让模型对文本重新排序。其目的是让模型学习文本中句子间的结构关系，以便更好的理解和生成连贯的自然语言文本。例如，在机器翻译、摘要生成等生成式任务中，模型需要生成具有流畅语境的自然语言文本，因此对句子的正确排序非常重要。

本文将训练样本中的每个文本语料分成多段文本，然后随机抽取50%的训练样本，将其中的多段文本随机打乱，让模型预测其顺序。模型首先需要理解每段文本中所包含的信息，包括每段文本的主题、语义和上下文信息。然后根据语境的先后顺序以及句子间的关联关系，给出

多段文本的顺序，将多段文本重新组合成具有流畅语境的自然文段。其损失值计算如公式（2）所示。

$$loss_{sort} = -\sum_{i=1}^{m} p_s \log p'_s \tag{2}$$

其中，p'_s为段落的顺序预测。p_s为真实答案。

3.1.4 句子间的距离关系

句子间的距离关系预测任务是预训练语言模型中的下一个句子预测任务的扩展，其目的是让模型能够学习文档级别的信息。文档级别的信息往往需要考虑到句子之间的关系，例如同一篇文章中的句子往往具有相关性，句子间的距离关系也能够体现文章的结构特点。因此该任务可以为其他自然语言处理任务提供有益的信息，例如文本分类、信息检索等任务。

该任务要求模型从一个文本序列中预测出句子间的距离关系，包括同一篇文章中相邻的句子、同一篇文章中不相邻的句子和不同文章中的句子。为了让模型学习到这些距离关系，本文采用了一种比较直观的方法：将给定的训练文本分成多段，随机选取一段文本，以25%的概率替换为同文档的其它句子，25%的概率替换为不同文档的句子，剩余50%的概率不替换，这种方法可以让模型在学习时考虑到不同文本之间的语境差异。

为了更好地训练模型，本文将句子间的距离关系任务视为三分类任务，其损失值计算如公式（3）所示。

$$loss_{relation} = -\sum_{i=1}^{3} p_r \log p'_r \tag{3}$$

其中，p'_r为句子间的距离关系预测。p_r为句子间的真实关系。

3.1.5 模型总体损失值

根据多任务学习框架，本文将上述各任务损失值的加权和作为模型总体损失值，如公式（4）所示。

$$loss_{all} = loss_{mlm} + \alpha loss_{sort} + \beta loss_{relation} \tag{4}$$

其中α和β是可调整的权重参数。

3.2 数据集

由于目前没有公开的大规模藏文知识语料库，为了增强模型的知识表示能力，本文构建了一个知识增强预训练数据集。本文通过爬取21个藏文网站如云藏网、西藏新闻网、西藏人民网等，收集大量的结构化知识和无结构化文本。然而，由于网络上的语料不够规范和完整，会存在大量的错误和噪声数据。因此，为了提高数据集的质量，在构建数据集的过程中，本文进行了以下数据清洗和预处理操作:

（1）本文将数据中的图片、链接、特殊字符等无意义的内容剔除。同时，过短的文本包含的文本信息不足，因此本文将词数低于100的文本去除，只保留了词数超过100的文本数据。

（2）爬取到的表格数据中偶尔会有不完整的三元组，如缺失实体或者关系，因此本文将信息不完整的三元组剔除。并且由于不同的文本语料中可能会包含相同的三元组，因此本文需要对三元组进行去重处理，确保知识库中每个三元组只出现一次。

通过以上的数据清洗和预处理，本文构建了一个包含50万个三元组、大小为4GB，藏文知识增强预训练数据集。该数据集共2.45亿个token。其中50万个三元组构成藏文知识库，无结构化文本作为藏文语料库。该数据集包含多个领域知识，如：经济、社会、科技、法律、体育等。最终用于增强预训练模型的知识表示能力。

第二十二届中国计算语言学大会论文集，第135页-第144页，哈尔滨，中国，2023年8月3日至5日。

3.3 词表构建

藏文是一种拼音文字，其单词的最小单位是一个音节，包含一个或最多七个字符，音节间以“.”来分割，但基于音节的分词并不能很好的表达语义结构。藏语的文字由一个或多个音节组成，同样以“.”分割。藏文包含七种结构：基字、上加字、下加字、前加字、后加字、后后加字和元音符号，共有155个字符。在预训练模型中，通常使用子词切分文本，而子单词表示可能不包含细粒度字符信息和全词的表示，本文同时编码藏文词表示和对应的藏文字符序列表示，使模型能够捕获不同粒度之间的语言知识，提升模型语言表达能力。

（1）字符表构建

除藏文的155个字符外，本文对训练语料中包含的其它字符进行了统计，频次在400以上的字符有941个，取其中前845个字符与藏文字符构成大小为1000的字符表，以做字符嵌入。

（2）子词表构建

如果以藏语文字进行分词，为了减少未被录入词表的单词数即未登录词（out-of-vocabulary，OOV），我们需要构建一个非常大的词表，这加大了机器的运算量，并且需要花费大量的时间和计算资源。针对OOV问题，本文使用sentencepiece(Kudo and Richardson, 2018)训练一个藏文分词模型，构建了一个大小为30,005，覆盖语料库99.99%字符的词表，并使用该分词模型对训练数据进行分词。

4 实验评估

本文使用文本分类、实体关系分类、机器阅读理解三个下游任务评估TiKEM模型性能。

4.1 藏文文本分类

本文藏语新闻数据集TNCC(Qun et al., 2017)，评估TiKEM模型对文本的分类能力。该数据集包含9,203条样本，涉及政治、经济、教育、旅游、环境、艺术、文学、宗教等12个领域。因原始数据集没有切分，本文按8:1:1的比例将其划分为训练集、验证集、测试集，评价指标为Accuracy和Macro-F1。

本文将知识增强预训练模型TiKEM与基于藏文音节分词的CNN分类模型、Transformer(Vaswani et al., 2017)、TextCNN(Guo et al., 2019)、DPCNN(Johnson and Zhang, 2017)等经典分类模型进行比较，同时也与少数民族多语言预训练模型CINO-base和藏文预训练模型TiBERT、BERT-base-Tibetan进行比较。实验结果如表1所示。

模型	Accuracy(%)	Macro-F1(%)
Transformer(Vaswani et al., 2017)	28.63	28.79
CNN(syllable)	61.51	57.34
TextCNN(Guo et al., 2019)	61.71	61.53
DPCNN(Johnson and Zhang, 2017)	62.91	61.17
TextRCNN(Lai et al., 2015)	63.67	62.81
BERT-base-Tibetan(安波and 龙从军, 2022)	-	51
TiBERT(Liu et al., 2022)	71.04	70.94
CINO-base(Yang et al., 2022)	73.1	70.0
TiKEM	**74.46**	**72.61**

表 1: 藏文文本分类结果

由表1可以看到，Transformer在藏文文本分类上的表现不如经典分类模型以及预训练语言模型，而TiKEM模型在藏文文本分类上的准确率超过了经典分类模型如TextCNN、DPCNN等，比TextCNN高了12.75%，DPCNN与TiKEM模型相差11.55%。同时TiKEM模型比TiBERT高了3.42%，并略微高于CINO-base模型。在Macro-F1值方面，TiKEM模型的表现同样超过了众多基线模型。与经典分类模型比较，CNN(syllable)与TiKEM模型之间相差15.27%，而TiKEM模型采用的同样是transformer结构，但是Macro-F1值却远远高于Transformer，这表明使用大规模藏文语料训练模型的方法提高了模型对藏文的理解能力。在预训练模型方面，TiKEM模型比BERT-base-Tibetan高

了21.61%，比CINO-base高了2.61%，比TiBERT高了1.67%，相对于只使用了大规模无结构化藏文语料训练的藏文预训练语言模型，融合藏文知识库的TiKEM模型在文本分类上有着明显的性能提升。

4.2 实体关系分类

为了验证TiKEM模型对知识的记忆及融合运用能力，本文构建了6,433条三元组-文本对齐数据集，三元组中共有11种关系。该任务要求在给定两个实体和包含该实体的对应文本后，给出两个实体之间的关系类别。本文按8:1:1的比例将其划分为训练集、验证集、测试集。本文使用FastText(Joulin et al., 2016)、DPCNN等作为基线模型，并与藏文预训练语言模型TiBERT和多语言预训练模型MiLMo、CINO-base进行比较。评价指标为Accuracy(%)、Macro-P(%)、Macro-R(%)和Macro-F1(%)，实验结果如表2所示。

模型	Accuracy(%)	Macro-P(%)	Macro-R(%)	Macro-F1(%)
FastText(Joulin et al., 2016)	55.80	34.05	32.98	31.61
DPCNN	70.94	54.21	49.23	48.65
TextCNN	72.38	71.03	59.11	56.76
TiBERT(Liu et al., 2022)	84.70	76.66	68.82	67.94
CINO-base(Yang et al., 2022)	85.31	75.48	69.12	66.73
MiLMo(Deng et al., 2023)	85.76	77.13	68.97	68.57
TiKEM	**90.12**	**91.73**	**75.61**	**76.34**

表 2: 藏文实体关系分类结果

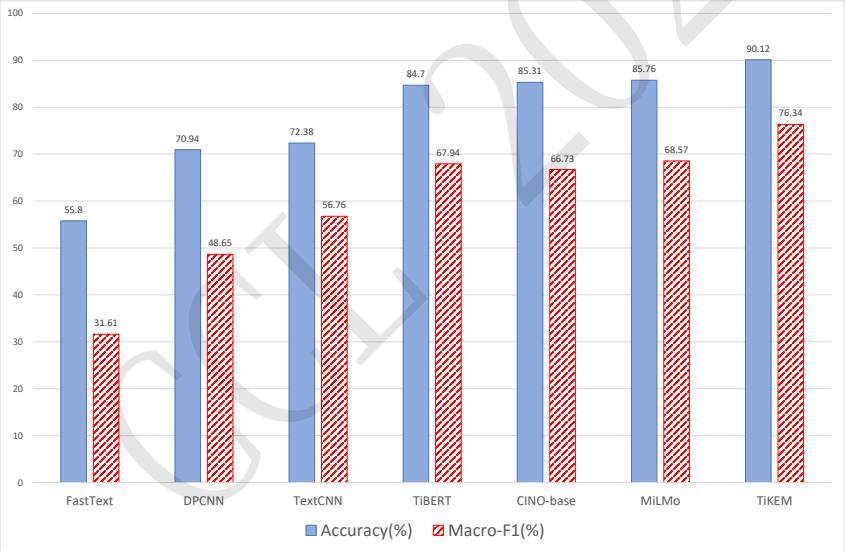

图 2: 模型在藏文实体关系分类中的Accuracy与Macro-F1值对比

由表2可以看到，FastText相对其它模型表现较差，TiKEM模型的准确率比FastText高了34.32%，比TextCNN高了17.74%。从总体上看，预训练模型在实体关系分类中比基线模型表现更好。同时融合了知识的藏文预训练模型TiKEM在该任务中，比TiBERT的准确率高了5.42%，比多语言预训练模型CINO-base和MiLMo的准确率分别高了4.81%和4.36%。

为了更清晰地观察各模型性能之间的差异，本文将各模型的准确率和Macro-F1值进行对比，绘制成柱状图，如图2所示。在以Macro-F1作为评价指标中，各模型的性能差异总体与准确率作为评价指标时的趋势一致。不同的是，在准确率中TiBERT比CINO-base低了0.61%。而在Macro-F1中，TiBERT比CINO-base高了1.21%。Macro-F1是各类别F1值的平均值，一定程度上反应了模型在不同类别中实体关系分类性能的偏差。因此可以看出CINO-base在一些类别中，实体关系分类性能比TiBERT更好，而总体上TiBERT的实体关系分类性能比CINO-base更稳定。此外，我们也可以看到，TiKEM模型在实体关系分类中的Macro-F1值高于其余模型。

其中，比预训练模型CINO-base、TiBERT和MiLMo分别高了9.61%、8.4%和7.77%。这表明，融合了知识的藏文预训练模型TiKEM有着更加丰富的知识，并且对给定的知识更加擅于去融合及运用。

4.3 藏文机器阅读理解

机器阅读理解任务是给定一段文本和一个问题，让模型回答对应问题。这需要模型理解问题和上下文语义，然后进行推理、判断等，给出具体答案。本文使用藏文机器阅读理解数据集TibetanQA(Sun et al., 2021c)对模型的阅读理解能力进行评估，该数据集包含了1,513篇文章和20,000个问答对。为了评估模型性能，本文使用EM值（精确匹配）和F1值作为评价指标。

本文以8:2的比例将数据划分为训练集和测试集，并使用机器阅读理解的经典模型R-Net(Wang et al., 2017)、BiDAF(Seo et al., 2017)、QANet(Yu et al., 2018)作为基线模型，这些模型在英文数据集上有着出色的表现，同时本文还将TiKEM与藏文预训练语言模型TiBERT和藏文抽取式机器阅读理解模型Ti-Reader(Sun et al., 2021b)进行比较。此外为了验证模型的知识推理能力，本文在TibetanQA数据集基础上，增加了1,823条包含三元组的藏文问答数据样本，并将数据同样以8:2的比例划分为训练集和测试集。实验结果如表3所示。

模型	TibetanQA		TibetanQA（含三元组）	
	EM(%)	F1(%)	EM(%)	F1(%)
R-Net(Wang et al., 2017)	55.8	63.4	-	-
BiDAF(Seo et al., 2017)	58.6	67.8	-	-
QANet(Yu et al., 2018)	57.1	66.9	-	-
TiBERT(Liu et al., 2022)	53.2	73.4	54.1	73.9
Ti-Reader(Sun et al., 2021b)	67.9	77.4	-	-
TiKEM	**69.4**	**80.1**	**72.6**	**81.3**

表 3: 预训练模型在藏文阅读理解上的应用

由表3可以看到，TiBERT在TibetanQA上F1值超过了R-Net等基线模型，但是EM值却低于基线模型。EM衡量模型预测与标准答案完全一致的占比，F1值评估模型预测与标准答案的重叠程度。F1值高而EM值低，说明TiBERT可以很好地确定答案范围，但是对于答案边界上的判断能力明显不足。而融合了知识之后的藏文知识增强预训练模型TiKEM在机器阅读理解上的性能相较于TiBERT有了极大的提升，并且EM值的提升幅度大于F1值的提升幅度。一方面实体掩码预测提高了模型对事物、人物、地点等实体的边界预测准确性，同时知识的融入提高了模型推理能力。另一方面句子关系预测等任务提高了模型对上下文结构和语境的理解能力。基于此，我们也可以看到TiKEM模型超过了藏文抽取式机器阅读理解模型Ti-Reader在TibetanQA上的表现。

在加入包含三元组的藏文问答数据样本后，TiBERT和TiKEM模型性能有了明显提升，TiBERT的EM值和F1值分别提升了0.9%和0.5%，TiKEM模型的EM值和F1值分别提升了3.2%和1.2%。显然，TiKEM模型提升幅度相较于TiBERT的提升幅度更大，这表明TiKEM模型比TiBERT更加擅长理解和运用知识，并进行知识推理。

5 总结

本文构建了一个包含50万个三元组的知识增强预训练数据集，在此基础上训练了一个基于知识增强的藏文预训练语言模型TiKEM，将结构化的藏文知识库和无结构化的文本统一表征。同时，针对知识的融合，将掩码预测任务扩展为实体掩码预测任务和知识掩码预测任务。为了学习句子间关系和文档级信息，本文将下一个句子预测任务扩展为句子重排序任务和句子间的距离关系任务。最后本文在在文本分类、关系分类、机器阅读理解三个下游任务上进行了实验。TiKEM模型性能均超过对比模型，证明了TiKEM模型在知识记忆、运用和推理能力等方面的有效提升。

第二十二届中国计算语言学大会论文集，第135页-第144页，哈尔滨，中国，2023年8月3日至5日。

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Junjie Deng, Hanru Shi, Xinhe Yu, Wugedele Bao, Yuan Sun, and Xiaobing Zhao. 2023. Milmo:minority multilingual pre-trained language model. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*.

Bao Guo, Chunxia Zhang, Junmin Liu, and Xiaoyi Ma. 2019. Improving text classification with weighted word embeddings via a multi-channel textcnn model. *Neurocomputing*, 363:366–374.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

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.

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.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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.

Wentao Ma, Yiming Cui, Chenglei Si, Ting Liu, Shijin Wang, and Guoping Hu. 2020. Charbert: Character-aware pre-trained language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 39–50.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Nuo Qun, Xing Li, Xipeng Qiu, and Xuanjing Huang. 2017. End-to-end neural text classification for tibetan. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 16th China National Conference, CCL 2017, and 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings 5*, pages 472–480. Springer.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*.

Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuan-Jing Huang, and Zheng Zhang. 2020. Colake: Contextualized language and knowledge embedding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3660–3670.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021a. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Yuan Sun, Chaofan Chen, Sisi Liu, and Xiaobing Zhao. 2021b. Ti-reader: 基于注意力机制的藏文机器阅读理解端到端网络模型(ti-reader: An end-to-end network model based on attention mechanisms for tibetan machine reading comprehension). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 219–228.

Yuan Sun, Sisi Liu, Chaofan Chen, Zhengcuo Dan, and Xiaobing Zhao. 2021c. 面向机器阅读理解的高质量藏语数据集构建(construction of high-quality tibetan dataset for machine reading comprehension). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 208–218.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

W Wang, N Yang, F Wei, B Chang, and M Zhou. 2017. R-net: Machine reading comprehension with self-matching networks. *Microsoft Research Asia, Beijing, China, Tech. Rep*, 5.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. 2021. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. Cino: A chinese minority pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949.

Adams Wei Yu, David Dohan, Quoc Le, Thang Luong, Rui Zhao, and Kai Chen. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *International Conference on Learning Representations*.

Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. 2021. Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9733–9740.

安波and 龙从军. 2022. 基于预训练语言模型的藏文文本分类. 中文信息学报, 36(12):85–93.

王海峰, 孙宇, and 吴华. 2022. 知识增强预训练模型. 中兴通讯技术, (16-24).