TiKG-30K：基于表示学习的藏语知识图谱数据集

庄文浩[1,3,&] 高歌[2,3,&] 孙媛[1,3,4,*]
[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，北京 100081
[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
[&]共同第一作者：庄文浩，高歌
*通讯作者：孙媛
tracy.yuan.sun@gmail.com

摘要

知识图谱的表示学习旨在通过将实体和关系映射到低维向量空间中来学习知识图谱数据之间的复杂语义关联，为信息检索、智能问答、知识推理等研究提供了支撑。目前知识图谱的表示学习研究主要集中在英、汉等语言，公开高质量数据集（如FB15k-237，WN18RR）对其研究起到非常重要的作用。但是，对于低资源语言（如藏语），由于缺少公开的知识图谱数据集，相关研究任务还处于起步阶段。基于此，本文提出一个公开的藏语知识图谱数据集TiKG-30K，包含了146,679个三元组，30,986个实体和641种关系，可应用于知识图谱的表示学习及下游任务。针对现有藏语知识图谱数据量少、数据稀疏的问题，本文利用藏文三元组中实体的同指关系，借助其他语言丰富的知识库和非文本介质对知识库进行扩充，通过跨语言近义词检索、合并同义实体和关系、修正错误三元组等技术对知识图谱进行多层优化，最终构建了藏语知识图谱数据集TiKG-30K。最后，本文采用多种经典表示学习模型在TiKG-30K进行了实验，并与英文数据集FB15k-237，WN18RR以及藏文数据集TD50K进行了对比，结果表明，TiKG-30K可以与FB15k-237、WN18RR数据集相媲美。本文将TiKG-30K数据集公开，https://tikg-30k.cmli-nlp.com/。

关键词： 藏语知识图谱；表示学习；知识图谱嵌入；链接预测

TiKG-30K: A Tibetan Knowledge Graph Dataset Based on Representation Learning

Wenhao Zhuang[1,3,&] Ge Gao[2,3,&] Yuan Sun[1,3,4,*]
[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
[&]Co-first authors: Wenhao Zhuang and Ge Gao
*Corresponding author: Yuan Sun
tracy.yuan.sun@gmail.com

Abstract

Representation learning of knowledge graphs aims to learn the complex semantic relationships between entities and relations in knowledge graph data by mapping them into a low-dimensional vector space, providing support for information retrieval, intelligent question answering, knowledge reasoning, and other research areas. Currently, research on representation learning of knowledge graphs mainly focuses on languages such as English and Chinese, and high-quality public datasets (such as FB15k-237, WN18RR) have played an important role in their research. However, for low-resource languages such as Tibetan, relevant research is still in the initial stages due to the lack of public knowledge graph datasets. In this paper, we propose a publicly available Tibetan knowledge graph dataset TiKG-30K, which contains 146,679 triples, 30,986

第二十二届中国计算语言学大会论文集，第145页-第154页，哈尔滨，中国，2023年8月3日至5日。

entities, and 641 relations, and can be applied to representation learning of knowledge graphs and downstream tasks. To address the problem of the small and sparse Tibetan knowledge graph dataset, we use the same-as relations between entities in Tibetan triples, and leverage other language-rich knowledge bases and non-text media to expand the knowledge base. We optimize the knowledge graph through multiple layers of techniques such as cross-lingual synonym retrieval, merging synonymous entities and relations, and correcting incorrect triples. Finally, we conduct experiments on TiKG-30K using classical representation learning models, and compare it with English datasets FB15k-237, WN18RR, and Tibetan dataset TD50K. The results show that TiKG-30K is comparable to FB15k-237 and WN18RR datasets. We make the TiKG-30K dataset public at https://tikg-30k.cmli-nlp.com/.

Keywords: Tibetan knowledge graph , Representation learning , Knowledge graph embedding , Link prediction

1 引言

藏语是中国少数民族语言之一，具有丰富的文化历史和独特的语言结构特点。近年来，随着人工智能领域的快速发展，基于藏语的知识图谱构建和知识表示学习成为研究热点之一。知识图谱是一种用于描述实体之间关系的结构化数据，它可以帮助我们更好地理解和利用丰富的实体知识。知识表示学习旨在将自然语言等符号化的知识转化为计算机可处理的向量表示(刘知远et al., 2016)，以便于机器学习算法的应用，表示学习的研究需要知识图谱作为数据支撑。

现有的知识图谱数据集大多针对中英文设计，如英文大规模通用知识图谱Freebase(Bollacker et al., 2008)，包含了超过8,000万个实体，三元组数量达到12亿条，从中提取知识事实构建的英文知识图谱FB15k(Bordes et al., 2013)，它包含14,951个实体、1,345个关系以及592,213个三元组。FB15k-237是FB15k的子集，在FB15k的测试集中，很多三元组可以通过训练集中简单的反转关系来获得，因此专家学者对出现的反向关系进行了去除，构建了更为有效的FB15k-237。中文大规模通用知识图谱CN-DBpedia(Xu et al., 2017)由复旦大学构建，涵盖超过2,200万的实体和2亿条三元组。相比之下，藏语等低资源语言的知识图谱比较匮乏，目前已有的藏语知识图谱如TD50K(Sun et al., 2021)，三元组数量为53,797，关系数量为3,285，三元组数据量较少，数据稀疏。针对现有藏语知识图谱数据量少、数据稀疏的问题，本文构建了一个藏语知识图谱数据集TiKG-30K，该数据集包含了30,986个实体以及641种关系类型，三元组数量为146,679，可以用于藏语表示学习的研究和链接预测、关系预测等相关任务。本文的主要贡献如下：

（1）针对现有藏语知识图谱数据量少、数据稀疏的问题，本文利用藏文三元组中实体的同指关系，借助其他语言丰富的知识库和非文本介质对知识库进行了扩充。

（2）在扩充三元组时，中英文专业词汇有时难以找到对应的藏语专业术语，导致产生歧义或者混淆语义。例如，中文学名“紫苏梗”、“紫草”、“紫花地丁”、“紫花针茅”对应着不同植物，但对应的藏语是相同的“ཟི་མིག་” （紫胶）。因此，本文采用三元组修正技术，合并同义实体和关系、删除不必要的实体和关系、修正错误的三元组等方式，进行了四个版本的优化更新，进一步构建了一个关系稠密、规模适中且适合用于表示学习任务的藏语知识图谱数据集TiKG-30K。

（3）采用TransE(Bordes et al., 2013)、DistMult(Yang et al., 2014)、ComplEx(Toutanova and Chen, 2015)、RotatE(Sun et al., 2019)、pRotatE(Sun et al., 2019)、HAKE(Zhang et al., 2020)多种经典表示学习模型在TiKG-30K进行了实验，并与英文数据集FB15k-237，WN18RR以及藏文数据集TD50K进行了对比，为藏文知识图谱表示学习提供了可开放测试的基线数据。

2 相关工作

知识图谱是谷歌在2012年提出的概念，其本质上是一种结构化的知识库，通过三元组（头实体，关系，尾实体）的形式来表示单条知识，以实体为节点，关系为边，大量的三元组可构建图谱结构，用来发掘不同实体间更为复杂的关系，现有的国外大规模通用知识图谱有DBpedia(Auer et al., 2007)、Yago(Suchanek et al., 2007)、Freebase(Bollacker et al., 2008)、Wikidata(Vrandečić and Krötzsch, 2014)等。国内大型知识图谱项目有CN-DBpedia(Xu et al., 2017)、XLORE(Wang et al., 2013)、openKG等，涵盖百科及细分领域的大量知识图谱。对于藏语知识图谱的构建，由于缺乏大规模的公开知识库，且三元组抽取技术的限制，现有的藏语知识图谱集中在特定领域，如汉藏双语旅游领域知识图谱(冯小兰and 赵小兵, 2019)，藏语数据集TD50K(Sun et al., 2021)等。此外，藏文知识图谱的数量与质量也远不能媲美中英文知识图谱。比如，在FB15k-237(Toutanova and Chen, 2015)英文数据集中，97.8%的实体出现两次及以上，每个实体平均拥有20个三元组，而在藏语数据集TD50K(Sun et al., 2021)中，只有48%的实体出现两次及以上，且一个实体平均仅有2个三元组，同时三元组数量也只有FB15k-237的17.3%。

知识图谱表示学习模型是一类广泛应用于知识图谱数据的机器学习模型。这些模型旨在通过将实体和关系映射到低维向量空间中来学习知识图谱数据之间的复杂语义关联，从而为许多基于知识图谱的任务提供更好的性能。2013年Bordes等人提出基于向量空间的表示学习模型TransE，它将实体和关系都映射到向量空间中，通过计算三元组中头实体与关系向量之和与尾实体向量的距离，来判断三元组是否成立(Bordes et al., 2013)。在TransE模型被提出后，TransH(Wang et al., 2014)、TransR(Lin et al., 2015)、TransD(Ji et al., 2015)等一系列基于TransE的改进模型被提出。2015年Yang等人提出的DistMult是一种基于张量分解的表示学习模型，它将实体和关系都表示为向量，并使用张量乘积来计算三元组的分数(Yang et al., 2014)。Toutanova等人提出一种基于复数向量的表示学习模型ComplEx，能够捕捉实体和关系之间的更复杂的交互(Toutanova and Chen, 2015)。RotatE是Sun等人在2019年提出的一种基于旋转操作的表示学习模型，它将关系表示为一个旋转矩阵，并通过对头实体和关系向量进行旋转，来预测尾实体，与前面几种模型相比，RotatE能够更好地处理对称性和反对称性关系(Sun et al., 2019)。2020年，在RotatE的工作基础上，Zhang等人对语义层次结构进行建模，将实体映射到极坐标系中，同心圆可以自然地反映等级，据此提出的HAKE模型在多个基准数据集上进行链接预测时取得了较好的结果(Zhang et al., 2020)，本文所用模型的简要信息如表1所示。

模型	得分函数	参数
TransE(Bordes et al., 2013)	$-\|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_{1/2}$	h, r, t $\in R^k$
DistMult(Yang et al., 2014)	$\mathrm{h}^\top diag(\mathbf{r})\mathbf{t}$	h, r, t $\in R^k$
ComplEx(Toutanova and Chen, 2015)	$\mathrm{Re}\left(\mathbf{h}^\top diag(\mathbf{r})\bar{\mathbf{t}}\right)$	h, r, t $\in C^k$
RotatE(Sun et al., 2019)	$-\|\mathbf{h}\circ\mathbf{r}-\mathbf{t}\|_2$	h, r, t $\in C^k, \|r_i\|=1$
HAKE(Zhang et al., 2020)	$-\ \|\mathbf{h}_m\circ\mathbf{r}_m-\mathbf{t}_m\|_2-\lambda$	$\mathrm{h}_m, \mathbf{t}_m \in R^k, \mathbf{r}_m \in R^k_+,$
	$\|\sin\left(\left(\mathbf{h}_p+\mathbf{r}_p-\mathbf{t}_p\right)/2\right)\|_1$	$\mathrm{h}_p, \mathbf{r}_p, \mathbf{t}_p \in [0,2\pi)^k, \lambda \in R$

表 1: 表示学习模型的得分函数与参数对比

3 TiKG-30K数据集的构建及优化

3.1 藏语知识图谱扩充

本文在前期爬取了中国藏族网通、宗喀巴网等许多藏文网站大量的原始藏文文章，并依据主题划分成了常识、旅游、法律、地理等不同的类型，使用词性标注系统对所有文章的字词进行了标注。参考高定国等人对藏语单句句型的研究(高定国and 扎西加, 2014)，本文首先对藏文文章中符合三元组提取的句型进行筛选，接下来，根据藏语单句中的词性标注，采用基于词性

第二十二届中国计算语言学大会论文集，第145页-第154页，哈尔滨，中国，2023年8月3日至5日。

组合的规则对单句中的结构进行拆分(扎西吉, 2018)，得到了主语、谓语、宾语等不同句子成分，组合成原始的三元组30余万条。

藏语知识库中三元组比较稀疏，因此如何借助其他语言丰富的知识库和非文本介质对知识库进行扩充，是本文的重要研究内容。我们通过藏文三元组中实体的同指关系（比如藏文三元组<སྐུ་འབུམ་དགོན།，རྒྱ་ཡིག་གི་མིང་།（中文名），塔尔寺>），借助于“塔尔寺”在百度百科、维基百科已有信息盒的“实体—属性—值”三元组关系，利用构建的18万汉藏对照词典，以及15,387条的藏汉语命名实体库，扩充藏文实体关系及属性值。另外，借助于地图兴趣点（POI）等非文本介质信息，获得对应中文实体的其他关系属性，传递给藏文实体，扩充藏文实体关系及属性值，解决藏语知识库数据量少、数据稀疏的问题。

对于低资源语言来说，单纯的获取文本语料并抽取实体和关系的规模比较有限，比如在藏语知识库中，存在<སྐུ་འབུམ་དགོན།，རྒྱ་ཡིག་གི་མིང་།（中文名），塔尔寺>三元组，而སྐུ་འབུམ་དགོན།（塔尔寺）这一实体只存在一个关系、属性三元组，这种数据稀疏的现象在藏语知识库中非常普遍。因此，本文借助其他语言丰富的知识库和非文本介质对知识库进行扩充，如图1所示。

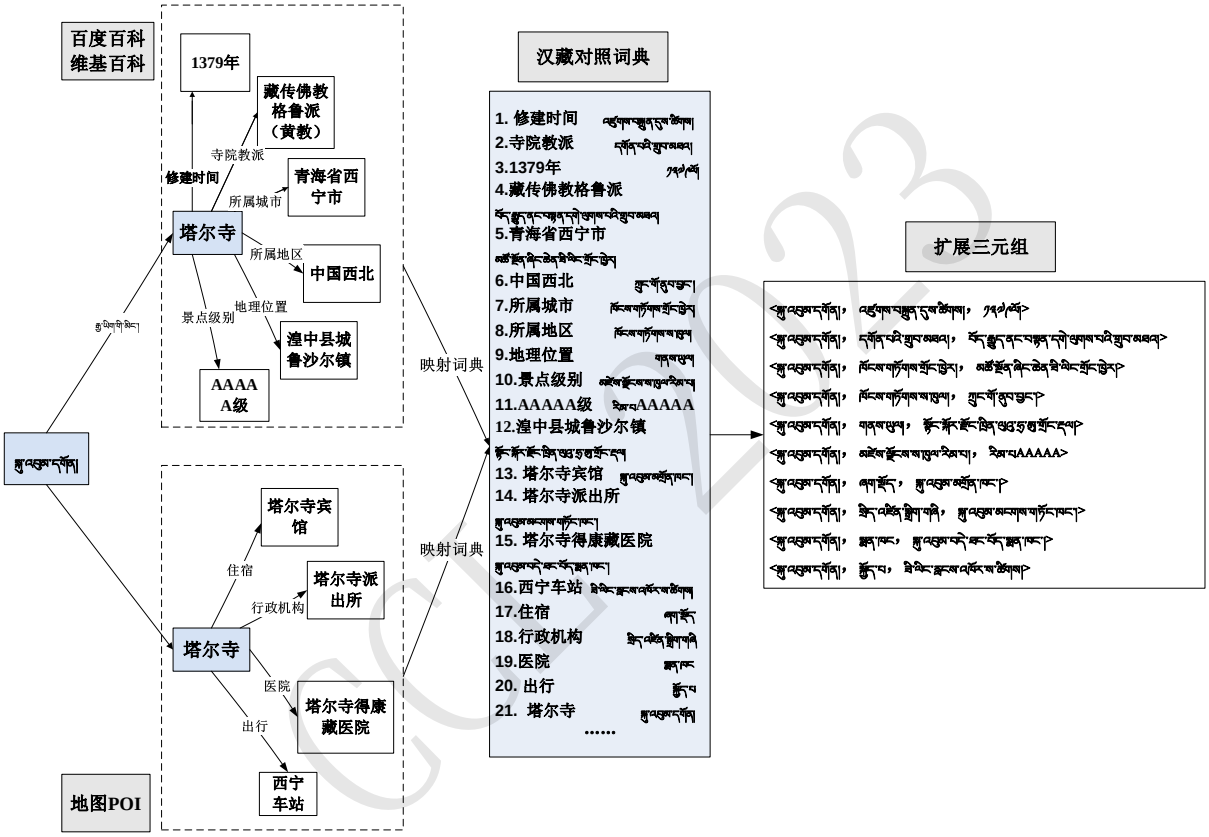

图 1: 藏语知识库扩充方法

我们借助百度百科、维基百科查找到“塔尔寺”这一实体，可以获得已有信息盒的“实体—属性—值”三元组关系，同时利用地图POI信息，扩充“塔尔寺”周边信息，最后利用汉藏对照词典，实现藏文实体关系属性的扩充，解决藏文知识库数据稀疏问题，如表2所示。结合百科与地图POI信息，我们将三元组数量扩充，去重后得到完整的三元组498,258条，此时实体数量达到348,596条，关系类型16,765种，以它作为初始的藏文知识图谱TiKG-V0。

3.2 藏语知识图谱优化

3.2.1 筛选实体和关系类型

TiKG-V1：在TiKG-V0中，由于实体和关系类型的数量相对较大，即使我们对藏语知识库进行了扩充，但整体依然很稀疏，不便于表示学习模型学习到很好的特征。因此，我们采用出现次数作为筛选标准，对所有实体和关系类型进行了统计。考虑到所需知识图谱的规模，我们

扩充前的藏文三元组	百度百科、维基百科、地图POI扩充	扩充后的藏文三元组
<སྐུ་འབུམ་དགོན།，ཀུ་ཡིག་གི་ཨིང་།，塔尔寺>	<塔尔寺，修建时间，1379年>	<སྐུ་འབུམ་དགོན།，འཛུགས་བསྒྲུབ་དུས་ཚོགས།，1379/ལོ།>
	<塔尔寺，寺院教派，藏传佛教格鲁派（黄教）>	<སྐུ་འབུམ་དགོན།，དགོན་པའི་གྲྭ་མཚན།，བོད་རྒྱུད་ནང་བསྒྲུབ་དགེ་ལུགས་པའི་གྲྭ་མཚན།>
	<塔尔寺，所属城市，青海省西宁市>	<སྐུ་འབུམ་དགོན།，ཞོང་གཏོགས་གོང་ཁྲེས།，མཚོ་སྨོན་ཞིང་ཆེན་ཞི་ཡིང་གོང་ཁྲེས།>
	<塔尔寺，所属地区，中国西北>	<སྐུ་འབུམ་དགོན།，ཞོང་གཏོགས་ས་ཁུལ།，ཀྲུང་གོ་རྒྱལ་བྱང་།>
	<塔尔寺，地理位置，湟中县城鲁沙尔镇>	<སྐུ་འབུམ་དགོན།，གནས་ཡུལ།，སྐོང་སྐོར་རྩོང་ཁྲིན་ལུའུ་ཏུ་ཚུ་ཏོང་དཀྱི།>
	<塔尔寺，景点级别，AAAAA级>	<སྐུ་འབུམ་དགོན།，མཐོ་སྡེད་ས་ཁུལ་རིམ་པ།，རིམ་པAAAAA>
	<塔尔寺，住宿，塔尔寺宾馆>	<སྐུ་འབུམ་དགོན།，ཞག་རྩོད་，སྐུ་འབུམ་མཇོན་ཁང་།>
	<塔尔寺，行政机构，塔尔寺派出所>	<སྐུ་འབུམ་དགོན།，ཕྱི་འཛིན་ལྷག་གཞི，སྐུ་འབུམ་མངགས་གཏོང་ཁང་།>
	<塔尔寺，医院，塔尔寺得康藏医院>	<སྐུ་འབུམ་དགོན།，སྨན་ཁང，སྐུ་འབུམ་བདེ་ཞང་བོད་སྨན་ཁང་།>
	<塔尔寺，出行，西宁车站>	<སྐུ་འབུམ་དགོན།，སྐྱོད་པ，ཟི་ཡིང་སྒངས་འཕོར་ཚོགས།>

表 2: 藏语知识库扩充

经过多次尝试，最终选择了筛选出现次数不小于8次的实体和不小于15次的关系类型。经过筛选，共得到166,086条三元组。

FB15k、WN18RR等常用于表示学习的数据集，往往需要有训练集（train）、验证集（valid）、测试集（test）三部分，且验证集和测试集中出现的实体和关系必须在训练集中出现过，这是因为这些数据集的目标是评估表示学习模型对于已知实体和关系的推理能力，而不是处理未知实体和关系。据此，我们再对166,086条三元组进行进一步的划分，首先选出127,664条三元组作为训练集，将剩下的三元组进行判定，若一条三元组的实体和关系都在训练集中出现过则保留，得到验证集和测试集各15,000条。

3.2.2 合并同义实体和关系类型

TiKG-V2：经过筛选后，TiKG-V1的稠密程度相较TiKG-V0有了很大提升，但是我们发现在TiKG-V1中依然存在一些问题，典型的就是语义重复的实体占据了一定数量。比如，“བེ་ཇིང་།”（北京）、“བེ་ཇིན་གོང་ཁྲེ་”（北京市）、“ཀྲུང་ཧྭ་མི་དམངས་སྤྱི་མཐུན་རྒྱལ་ཁབ་ཀྱི་རྒྱལ་སྡ་”（中华人民共和国首都）、“ཀྲུང་གོའི་བེ་ཇིན།”（中国北京），这些实际意义相同的实体在数据集中以独立实体出现，倘若能将这些实体合并，会降低数据冗余，构建出关系更丰富的图谱，无论对于知识推理还是提高查询准确率都有一定帮助。

关于如何寻找藏语近义词的研究较少，基于规则的寻找方法也不够准确。考虑到在跨语言对比中，有些近义词往往会有相同的映射，如果将同一映射中的内容粗略地视作近义词，会合并很多近义实体和关系。采用这种方式，本文对符合条件的近义词进行合并，合并之后会出现少许重复，再次去重得到TiKG-V2。

TiKG-V3：在TiKG-V2中，利用跨语言对比的方式虽然能合并一些近义实体，但仍然会漏掉一些，比如上述的“བེ་ཇིང་།”（北京）和“བེ་ཇིན་གོང་ཁྲེ་”（北京市）在跨语言对比中容易得到相同的映射，但“ཀྲུང་ཧྭ་མི་དམངས་སྤྱི་མཐུན་རྒྱལ་ཁབ་ཀྱི་རྒྱལ་སྡ་”（中华人民共和国首都）仅仅通过跨语言对比很难与前两个实体进行合并。

通过更细致的观察，我们发现在扩充的三元组中，很多头实体并不相同，但是对应的关系和尾实体相同，这是因为在扩充时，很多别名或者代称在百科中会有相同的搜索结果，例如，“ཤར་ཧོག་ཀྱི་རྒྱ་ཆེ་ཁ།”（东方之珠）、“དངོས་ཟོག་ཚོང་ཁྱབ་བྱེད་ས།”（购物天堂）、“གྲུང་གིའི་ཞང་ཀང་།”（中国香港）这三个关键词在百科中都会对应到香港特别行政区。对此，我们在TiKG-V1的基础上不进行TiKG-V2的合并，直接对有着相同关系类型及尾实体的头实体进行合并，并且把TiKG-V2中有相同映射的联系进行合并，得到TiKG-V3。

TiKG-30K：在以上几个版本的优化中，仅靠跨语言对比、寻找等方式并不能充分、正确地对近义词进行合并，因此在最后，需要加入人工来审查及合并。在TiKG-V3的基础上，我们检查了关系类型，调整不准确的近义关系，合并了新的关系类型。检查已经合并的实体，修正少量错误，同时参考TiKG-V2中具有相同映射的近义实体，对语义相同但格式有区别的实体（如“གྲུང་གོ་ཅོ་ནན་ནན་དཔང་།”（中国·河南·南阳）、“ཅོ་ནན་ཞིང་ཆེན་ནན་དཔང་ཚོང་ཁྲོང་།”（河南省南阳市））、尾实体中与头实体语义相似的实体进行合并。

在完成上述人工审查合并工作后，我们得到了最终的知识图谱TiKG-30K。它的实体和关系类型较为常见，内容与中国地理具有较强相关性，实体和关系可靠准确，数据公开，开放性和可扩展性强，能够为针对藏语知识图谱的表示学习模型、藏语问答系统等领域提供支持。TiKG知识图谱各版本信息如表3所示。

数据集	实体	关系类型	训练集	验证集	测试集
TiKG-V0	348,596	16,765	498,258	-	-
TiKG-V1	34,836	698	127,664	15,000	15,000
TiKG-V2	33,521	659	125,995	14,960	14,958
TiKG-V3	32,164	659	117,633	14,834	14,818
TiKG-30K	**30,986**	**641**	**117,051**	**14,820**	**14,808**

表 3: TiKG知识图谱各版本的对比

4 实验结果与分析

在一般情况下，数据集关系类型的数量远少于实体数量，所以实体链接预测难度相较关系预测难度也更大(曾平, 2018)，因此本文在测评任务中选择难度较大的实体链接预测。

4.1 评估指标

（1）Mean Reciprocal Rank：平均倒数排名，简称MRR，用于衡量知识图谱中实体链接任务性能的常用指标之一。给定一个查询q_i，模型需要在知识图谱中为其找到相应的实体。由于知识图谱中可能存在多个与查询实体相似的实体，因此需要对这些实体被选中的概率值由高到低排序，假设正确的尾实体e_i的排名为rank(e_i)。对所有的链接预测任务计算正确排名的平均值，可以得到Mean Rank（平均排名）指标，简称MR，但是该指标受待预测实体数量的影响，并不能真实地反映模型的性能。

$$MR = \frac{\sum_{q_i \in Test} rank(e_i)}{\mid Test \mid}$$

而MRR将排名的倒数进行求和取平均，将评测指标范围限制在0 ∼ 1之间，数值越大，客观上说明模型的性能越好，因此本文采用MRR作为评测指标之一。

$$MRR = \frac{\sum_{q_i \in Test} \frac{1}{rank(e_i)}}{|Test|}$$

（2）Hits@k：k命中率，正确排名rank(e_i)排在top-k所占的比例，取值范围0 ∼ 1，取值越大，模型性能越好，在本文测评中，k值分别取1，3，10。另外需要注意的是，给定头实体和关系进行预测时，在有多个正确的尾实体时，如果测试样例给出尾实体的排名为n，排在前边n-1个的实体中如果出现正确的尾实体也会被模型认为是错误的，这会影响正确的实验结果。因此在进行评测时，需要先将其他正确的尾实体过滤掉，经过过滤处理的记为Filtered，未经

处理的记为Raw。因过滤处理后的评测结果更为合理，故本文的评测指标全部基于过滤后的数据。

$$Hits@k = \frac{\sum_{q_i \in Test} I\left(\frac{1}{rank(e_i)} \leq k\right)}{|Test|}$$

4.2 TiKG-30K与基准数据集对比实验结果

由于目前公开可用的藏文数据集相对较少，我们使用TD50K(Sun et al., 2021)作为藏文基准数据集。TiKG-30K与基准数据集的对比如表4所示。

数据集	实体	关系类型	训练集	验证集	测试集
WN18RR	40,943	11	86,835	3,034	3,134
FB15k-237	14,541	237	272,115	17,535	20,466
TD50K	12,573	3,285	27,754	9,253	9,251
TiKG-30K	30,986	641	117,051	14,820	14,808

表 4: TiKG-30K与基准数据集的对比

与已有的藏文知识图谱TD50K相比，TiKG-30K提供更多、更全面、更准确的实体关系信息，模型能够学到更好的表示，根据表5实验结果，TiKG-30K更适合用于表示学习的研究。

	TransE				DistMult			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TD50K	-	-	-	.25	-	-	-	.31
TiKG-30K	.496	.419	.548	**.625**	.399	.367	.416	**.457**

表 5: TiKG-30K与TD50K的实验对比结果

使用链接预测常用基准数据集FB15k-237、WN18RR与TiKG-30K进行对比实验。WN18RR是WN18(Toutanova and Chen, 2015)数据集的子集，由于WN18中存在测试集泄露的问题，因此改进的WN18RR内容更加合理。对比实验结果如表6所示。

	TransE				DistMult			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
WN18RR	.226	-	-	.501	**.43**	**.39**	**.44**	**.49**
FB15k-237	.294	-	-	.465	.241	.155	.263	.419
TiKG-30K	**.496**	.419	.548	**.625**	.399	.367	.416	.457
	ComplEx				RotatE			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
WN18RR	.44	.41	.46	.51	.476	.428	.492	.571
FB15k-237	.247	.158	.275	.428	.338	.241	.375	.533
TiKG-30K	**.479**	**.437**	**.502**	**.554**	**.529**	**.483**	**.553**	**.612**
	pRotatE				HAKE			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
WN18RR	.462	.417	.479	.552	.497	.452	.516	.582
FB15k-237	.328	.230	.365	.524	.346	.250	.381	.542
TiKG-30K	**.526**	**.468**	**.557**	**.630**	**.534**	**.483**	**.561**	**.629**

表 6: TiKG-30K与英文基准数据集的实验对比结果

将TiKG-30K与WN18RR、FB15k-237的Hits@10指标进行直观对比，如图2所示。结合表5、6进行分析，本文提出的TiKG-30K在实验中的各项指标相较基准数据集均有所提高，说明TiKG-30K在知识图谱链接预测任务上具有更好的性能表现。

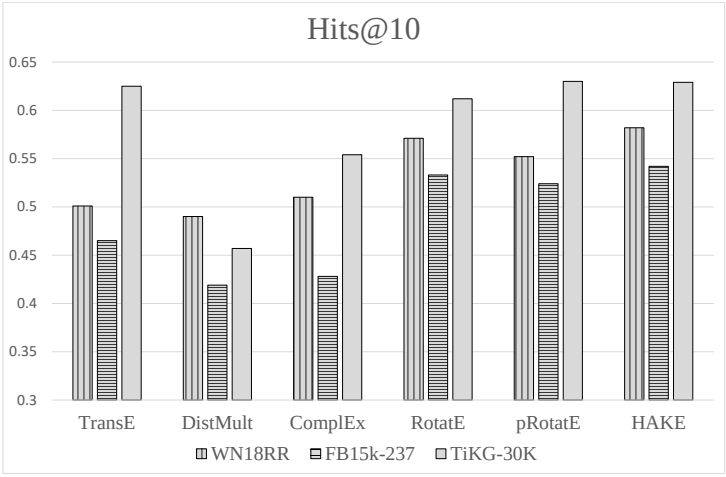

图 2: TiKG-30K与WN18RR、FB15k-237在不同模型上Hits@10的实验结果

4.3 消融实验结果

本文构建TiKG-30K时，通过跨语言近义词检索、合并同义实体和关系、修正错误三元组等技术对知识图谱进行多层优化，为了验证优化方式有效，需要对TiKG-V1、TiKG-V2、TiKG-V3、TiKG-30K四个不断优化的知识图谱进行消融实验。对于每个进行链接预测的模型，在相同的参数设置下，分别记录MRR、Hits@1、Hits@3、Hits@10指标的实验结果如表7所示，将实验结果中Hits@10指标进行对比如图3所示。

	TransE				DistMult			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TiKG-V1	.440	.346	.505	.594	.360	.322	.382	.428
TiKG-V2	.446	.351	.514	.599	.358	.320	.381	.426
TiKG-V3	**.496**	**.423**	.542	.621	**.403**	**.371**	**.422**	**.460**
TiKG-30K	**.496**	.419	**.548**	**.625**	.399	.367	.416	.457
	ComplEx				RotatE			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TiKG-V1	.432	.383	.462	.521	.484	.430	.512	.578
TiKG-V2	.436	.388	.465	.521	.492	.441	.519	.584
TiKG-V3	.476	.435	.499	.551	.528	**.485**	.549	.609
TiKG-30K	**.479**	**.437**	**.502**	**.554**	**.529**	.483	**.553**	**.612**
	pRotatE				HAKE			
	MRR	Hits@1	Hits@3	Hits@10	MRR	Hits@1	Hits@3	Hits@10
TiKG-V1	.480	.410	.523	.604	.490	.427	.526	.601
TiKG-V2	.490	.423	.529	.608	.498	.437	.533	.604
TiKG-V3	.520	.464	.550	.623	.528	.477	.554	.622
TiKG-30K	**.526**	**.468**	**.557**	**.630**	**.534**	**.483**	**.561**	**.629**

表 7: TiKG各版本数据集的消融实验对比结果

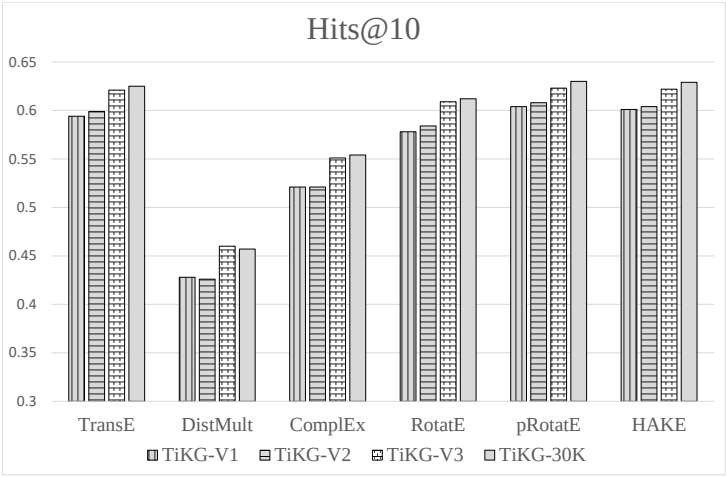

图 3: 不同模型上Hits@10的消融实验结果

综合上述消融实验结果，可以得出以下结论：

（1）在ComplEx、pRotatE、HAKE模型上进行链接预测任务时，TiKG-30K的各项指标均领先于前三个版本数据集的同类指标；

（2）在TransE、RotatE模型中，TiKG-30K仅有Hits@1指标落后于TiKG-V3，平均落后0.3%；仅在DistMult模型中，TiKG-30K各项指标小幅度落后于TiKG-V3；

（3）总体上TiKG-V1、TiKG-V2、TiKG-V3、TiKG-30K在不同模型中的各项指标均随着数据集的优化而提升，这证实本文的优化方式是有效的。综上可以认为TiKG-30K是这四个版本中结构更加合理、内容更加准确的知识图谱数据集，更适合应用于藏语表示学习领域。

5 总结与展望

本文构建了藏语知识图谱TiKG-30K用于藏语表示学习的研究，借助百科与地图POI信息对知识图谱进行扩充，有效缓解了藏文知识图谱三元组数量少、数据稀疏的问题，通过多种方式合并近义实体和关系，减少数据冗余和语义重复，有效解决了知识图谱中关系稀疏的问题。在TransE、DistMult、ComplEx、RotatE、pRotatE、HAKE表示学习模型上进行链接预测任务评估，实验结果表明，对比基准数据集与TiKG各版本数据集，TiKG-30K有着更好的表现，证实本文数据优化方式的有效性。由于TiKG-30K比已有的藏语知识图谱有着更丰富合理的实体关系表示，所以它能够对藏语表示学习的研究、知识推理以及藏语问答系统等多个领域提供高质量的基础数据。在未来我们将探索更多数据源，对现有的藏语知识图谱进行扩充，对TiKG-30K进行更多方向的优化，以能够胜任更多藏语自然语言处理的任务，推动自然语言处理在藏语等低资源语言领域的发展。

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（22ZD035）的资助。

参考文献

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007+ ASWC 2007, Busan, Korea, November 11-15, 2007. Proceedings*, pages 722–735. Springer.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, pages 687–696.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197*.

Yuan Sun, Andong Chen, Chaofan Chen, Tianci Xia, and Xiaobing Zhao. 2021. A joint model for representation learning of tibetan knowledge graph based on encyclopedia. *Transactions on Asian and Low-Resource Language Information Processing*, 20(2):1–17.

Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*, pages 57–66.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Zhigang Wang, Juanzi Li, Zhichun Wang, Shuangjie Li, Mingyang Li, Dongsheng Zhang, Yao Shi, Yongbin Liu, Peng Zhang, and Jie Tang. 2013. Xlore: A large-scale english-chinese bilingual knowledge graph. In *ISWC (Posters & Demos)*, pages 121–124.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28.

Bo Xu, Yong Xu, Jiaqing Liang, Chenhao Xie, Bin Liang, Wanyun Cui, and Yanghua Xiao. 2017. Cn-dbpedia: A never-ending chinese knowledge extraction system. In *Advances in Artificial Intelligence: From Theory to Practice: 30th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2017, Arras, France, June 27-30, 2017, Proceedings, Part II*, pages 428–438. Springer.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*.

Zhanqiu Zhang, Jianyu Cai, Yongdong Zhang, and Jie Wang. 2020. Learning hierarchy-aware knowledge graph embeddings for link prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 3065–3072.

冯小兰and 赵小兵. 2019. 汉藏双语旅游领域知识图谱系统构建. 中文信息学报, 33(11):64–72.

刘知远, 孙茂松, 林衍凯, and 谢若冰. 2016. 知识表示学习研究进展. 计算机研究与发展, 53(2):247–261.

扎西吉. 2018. 基于pcfg的藏语句法分析. Master's thesis, 青海师范大学.

曾平. 2018. 基于文本特征学习的知识图谱构建技术研究. Ph.D. thesis, 国防科技大学.

高定国and 扎西加. 2014. 藏语单句的基本句型研究. 中国藏学, (4):127–133.