# 数字人文视域下的青藏高原文旅知识图谱构建研究 ——以塔尔寺为例

**李鑫豪　　赵维纳　　赵婉亦　　李超群**
青海师范大学计算机学院，西宁810001
1020603546@qq.com　　490333294@qq.com　　337081897@qq.com
lcq513@163.com

## 摘要

青藏地区多元的民族构成以及悠久的历史沉淀孕育出丰富且独特的青藏文化，使得这片雪域圣地焉然成为了“高原文化宝库”。然而受闭塞的交通条件和较滞后的经济水平的限制，青藏地区文旅资源的保护与弘扬工作始终处于滞后状态。本文以数字人文为导向，在提示学习框架下采用联合学习的方式对文本中实体与关系的抽取，实现低资源条件下的知识抽取，形成一套文旅知识图谱构建范式，并以全国重点文物保护单位‘塔尔寺’为代表，完整的介绍了塔尔寺知识图谱从本体设计、原始数据获取、知识抽取到可视化展示的详细流程。最终，本文所构建的塔尔寺知识图谱共包含4705个节点及17386条关系。本文的工作弥补了人文领域青藏文化的结构化数据不足的问题，同时为青藏文旅在数字人文领域的研究提供参考。

**关键词：** 青藏文化；提示学习；联合抽取；塔尔寺知识图谱

# Research on the Construction of Cultural and Tourism Knowledge Atlas on the Qinghai-Tibet Plateau from the Perspective of Digital Humanity ——A case study of Kumbum Monastery

**Xinhao Li　　Weina Zhao　　Wanyi Zhao　　Chaoqun Li**
School of computingQinghai Normal University，Xining810001
1020603546@qq.com　　490333294@qq.com　　337081897@qq.com
lcq513@163.com

## Abstract

The diverse ethnic composition and long history of the Qinghai-Tibet region have bred a rich and unique Qinghai-Tibetan culture. Making this snowy sacred place a treasure trove of plateau culture ”. However, due to blocked traffic conditions and lagging economic level, the protection and promotion of cultural and travel resources in Qinghai-Tibet region is always lagging behind. This paper, guided by digital humanities, implements the extraction of entities and relationships in text by means of joint learning under the prompting learning framework, achieves the extraction of knowledge from low resources, forms a set of cultural and travel knowledge graph construction paradigm, and takes the national key cultural relics protection unit ’Kumbum Monastery’ As a representative, the detailed process from ontology design, original data acquisition, knowledge extraction to visual display of the knowledge graph of Kumbum Monastery is introduced. Finally, the knowledge map of Tar Temple built in this paper contains 4705 nodes and 17386 relations. The work of this paper makes up for the shortage of structured data of Tibetan culture in the field of human culture, and provides reference for the study of Tibetan travel in the field of digital human culture.

第二十二届中国计算语言学大会论文集，第253页-第263页，哈尔滨，中国，2023年8月3日至5日。

**Keywords:** Qinghai-Tibet Culture , Prompt learning , joint extraction , Kumbum Monastery knowledge graph

## 1 引言

青藏地区拥有着多元的民族构成以及悠久的历史沉淀，定居于青藏高原的50余个民族之间产生的文化交融与碰撞，促成了青藏地区人文历史的发展演变。“丝绸之路”南线的青海道、唐蕃古道、茶马古道(刘峰贵et al., 2012)留存于河湟谷地，其沿线分布着素有“世界屋脊之珠”的布达拉宫，藏传佛教圣地塔尔寺，被誉为“海藏咽喉，茶马商都”的丹噶尔古城以及“藏式建筑的千古典范”的大昭寺。除此之外青藏高原还拥有大量的非物质文化遗产，藏戏、藏医、热贡艺术这些弥足珍贵的遗产对中华传统文化产生着长期、广泛而又深刻的影响。这片辽阔的土地所孕育出的风格独特、形式多样的民族文化、历史文化、宗教文化和民俗文化焉然构成了“高原文化宝库”。然而受闭塞的交通条件与欠发达的经济水平的限制，青藏地区的文化保护与弘扬工作处于较为滞后状态。如何开展青藏文化的保护与传承工作是引人深思的。

数字人文是将现代化数字技术融入人文学科研究的一种典型的文理交叉领域学科(翁冉et al., 2023)，具体而言就是将已有的大量数字化资料，借助计算机技术的辅助分析，从海量数据中发现隐藏在数据中的模式、知识和趋势，揭示事物发生与发展的规律，对未来发展进行预测。随着数字人文研究愈来愈受关注，青藏高原的文旅产业也迎来了新的机遇。

知识图谱（Knowledge Graph）是一种新的知识组织方式，它采用（实体，关系，实体）的三元组形式进行表达和组织(Paulheim and Heiko, 2017)，如同蛛网一般链接起海量的异构信息，得易于其天然的图结构，知识图谱在查询与推理方面拥有显著优势。当前，知识图谱在文化领域得到了初步应用，在传统文化保护与传承、旅游推广等领域发挥着越来越重要的作用。如(聂欣晗and 张亮玉, 2021)利用知识图谱对《红楼梦》、《三国演义》等经典著作进行知识组织、分析与推理，在数字人文视觉下实现对传统文化的深入挖掘。可以看出，知识图谱在文化领域有重要的应用前景。然而知识图谱在青藏文旅领域的应用还较少，本文初步探索以知识图谱为代表的数字人文技术在青藏高原文旅中的应用。

具体地，本文提出一套文旅知识图谱构建范例，并在提示学习框架下利用联合学习实现实体与关系的抽取，探索低资源情况下知识的抽取。同时，在预训练语言模型的基础上进行链接，实现不同数据源的知识融合。实验结果表明，该方法能够在少量标注数据的情况下，实现高质量的知识抽取及融合，形成文旅知识图谱。最后，利用Neo4j数据库实现了知识图谱的存储与展示。本工作的意义在于为青藏文化保护工作的开展提供了新模式，弥补了青藏文旅数据的结构化不足的问题，从数字人文的角度实现对传统文化的传承与保护。

## 2 相关工作

### 2.1 数字人文

近年来,如何利用数字化技术激发创新创造活力，推动文化旅游产业高质量发展已成为一项重要课题，研究人员对数字人文进行了不同领域及层次的研究，(孙乃荣et al., 2022)以河北省文化和旅游外宣网站切入点，针对当前网站存在的问题，提出相应解决方法，从而助力河北文旅产业的发展。(刘泽权, 2021)通过重新界定名著重译的概念、对其评价方法进行梳理与审视，并从数字人文视角出发，提出了名著重译等级评价模型及相关变量，最后以《老人与海》、《红楼梦》和《香菱学诗》为例展示了模型的操作流程。(张卫et al., 2021)以古诗文本为例，利用机器学习与深度学习实现面向汉语诗文及其鉴赏的大规模人文情感术语的自动化抽取与分析。并得出将现代鉴赏融入古诗原文可显著优化情感知识的广度与深度的结论。


**基金项目**：省部共建藏语智能信息处理及应用国家重点实验室自主课题基金项目面向藏文古文献知识图谱构建（2022-SKL-012）国家自然基金项目汉藏双语藏医药知识图谱构建技术研究（62266036）国家社科基金项目藏汉双语藏文古籍知识图谱构建研究（22BTQ010）
**通讯作者**：赵维纳

### 2.2 文旅知识图谱

文旅领域知识图谱的研究同样备受关注。旅游方面，(刘济源, 2019)提出一种改进的本体构建方法利用信息抽取与知识融合技术从多源异构的数据中构建出旅游知识图谱，以此为基础设计出旅游问答系统，进一步改进设计出智能旅行助手。(张宇飞et al., 2022)利用爬虫技术从网络中获取河北省旅游景区的原始数据，将景区的信息进行分析、整合，构建出河北省旅游景区知识图谱，从而解决了河北省景点数量多分布广、信息杂糅且线上搜索智能化不足的问题。文化方面，(周莉娜et al., 2019)通过调研唐诗领域知识服务需求，设计出唐诗本体模型，对唐诗领域海量数据的语义化处理和存储构建唐诗知识图谱，基于该图谱搭建了面向领域知识服务的唐诗智能服务平台。虽然文旅领域知识图谱发展迅猛，但当前涉及青藏文化领域的知识图谱鲜有开展，随着中华文化保护与弘扬的提出，构建面向青藏文化领域的知识图谱成为一件具有重要意义的工作。

### 2.3 知识抽取

知识抽取是从半结构化或非结构化的文本中，抽取出机器可以理解和处理的知识，它主要包含实体识别、关系抽取、事件抽取三个子任务。基于深度学习方法的知识抽取,是近年的研究热点,其能够将低层特征进行组合,形成更加抽象的高层特征。

基于深度学习知识抽取早期通常采用流水式抽取法，即实体识别与关系抽取互相独立进行。虽然流水式抽取不断优化迭代，但其始终面临着两个子任务间存在错误传播、无关联实体间产生冗余信息以及信息丢失的问题。研究人员也因此开始尝试将实体与关系进行联合抽取。

联合抽取方法是通过实体识别和关系分类联合模型,直接得到存在关系的实体三元组。(Miwa and Bansal, 2016) 首次提出了基于深度学习的联合抽取模型，其利用双向顺序结构和双向树形结构的LSTM-RNNs进行实体和关系的联合建模。虽然该模型中的关系分类子任务和实体识别子任务仅共享了编码层的双向序列LSTM表示，但它的提出为日后真正意义上基于参数共享的联合学习奠定了基础。(Zheng et al., 2017)提出了基于新的标注策略的实体关系抽取方法,将原来涉及到命名实体识别和关系分类两个子任务的联合学习模型完全变成了一个序列标注问题。(Wang et al., 2020)提出了一种名为TPLinker（Token Pair Linking）的实体关系联合抽取标注方案，该方案可在一个模型中实现单阶段联合抽取，模型不存在曝光偏差，保证训练和测试的一致性。并且同时可解决多关系重叠和多关系实体嵌套的问题。当前基于深度学习的联合抽取方式已经成为了关系抽取的主流方式。

## 3 数字人文视域下的青藏高原文旅知识图谱构建框架

青藏高原文旅类知识图谱的构建包括：本体构建、数据获取、模型训练、知识抽取、知识融合等流程。以下将以塔尔寺知识图谱构建为范例，分别按照步骤介绍构建流程：

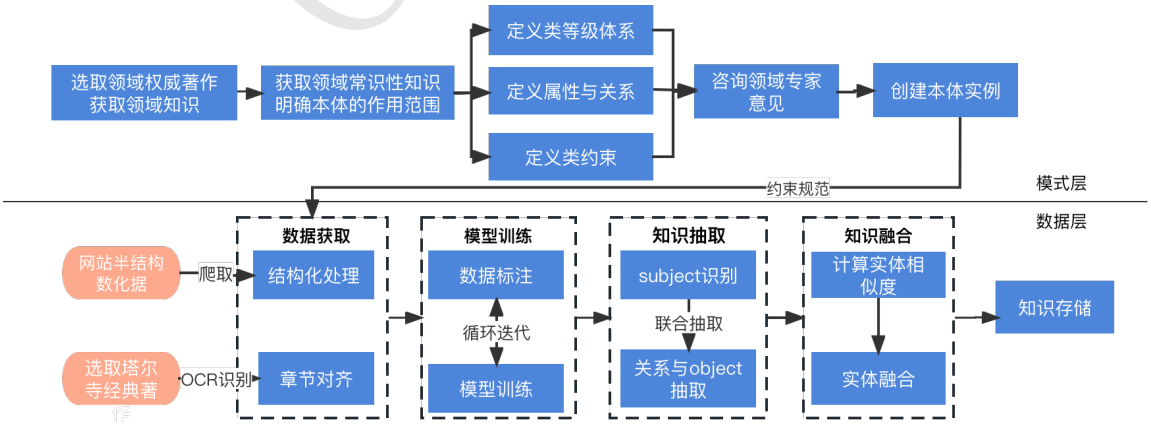


图 1: 青藏高原文旅知识图谱构建流程

### 3.1 数据源获取

本文以塔尔寺相关著作为数据源，利用OCR工具将《塔尔寺史话》、《圣地莲花塔尔寺》、《塔尔寺建筑艺术史》等著作的纸质文本转化为数字文本，并将不同著作中内容相关的章节进行章节整合，最后选取每个章节的前百分之二十进行标注，标注过程与训练过程迭代进行，模型的训练结果作为反馈对下一轮次的标注进行指导。为保障所构建的原始语料库尽可能完善，本文还利用爬虫技术在青海省图书馆塔尔寺知识平台和青海文化旅游厅爬取了有关青海省文化遗产的结构化数据，最终得到塔尔寺原始语料库。

### 3.2 本体设计

本体是对实体存在形式的描述，往往表示为一组概念定义和概念之间的层级关系，本体构建是知识图谱的构建的基础工作，本文将结合七步法(Abdellatif et al., 2017)与骨架法(傅柱et al., 2013)提出一种创新性的本体构建方式，为日后青藏文化领域相关知识图谱的模式层构建提供参考。

（1）选取要构建领域的权威著作或文献；通读著作，尤其需要针对著作内所存在的命名实体保持敏感，当一类实体频繁出现时需及时记录实体类型及其属性。以塔尔寺图谱构建过程所选取的《塔尔寺史话》为例，书中频繁出现人物、寺院、大师、佛像、法会等实体类型。

（2）获取领域常识性知识；若在通读著作的过程中遇到领域内常识性问题，需要适当的加以学习，这些领域内常识性问题往往对明确实体类型的边界，以及后续制定标注规则有着重要作用。同样以塔尔寺图谱构建过程所选取的《塔尔寺史话》为例，书中并未介绍大师、喇嘛、班禅以及活佛的区别，但这些实体类型对人物实体的层级划分以及后续标注规则的制定着重要影响。

（3）定义类及其等级体系、类属性、类约束等；通过先前的两个步骤，已经能够大致确定图谱的实体类型及其属性，在定义类的层级划分时可参考著作目录，不同类型的实体在目录中大致都会得到体现。塔尔寺知识图谱的构建过程中共定义实体31类，关系20种。

（4）咨询领域专家建议；本体构建作为图谱构建的基础工作，后续改动所造成的额外工作量十分庞大，为确保后续工作顺利进行，可向领域专家说明任务需求，咨询专家建议，对构建出的类以及关系体系进行修改。

（5）创建实例；在创建本体实例时可以利用Protégé软件，该软件拥有便捷的实体、关系以及属性的管理功能，本体实例创建完成后可进行界面化展示，直观的展示出各实体类型之间的关系。如图2展示了塔尔寺知识图谱的本体结构。

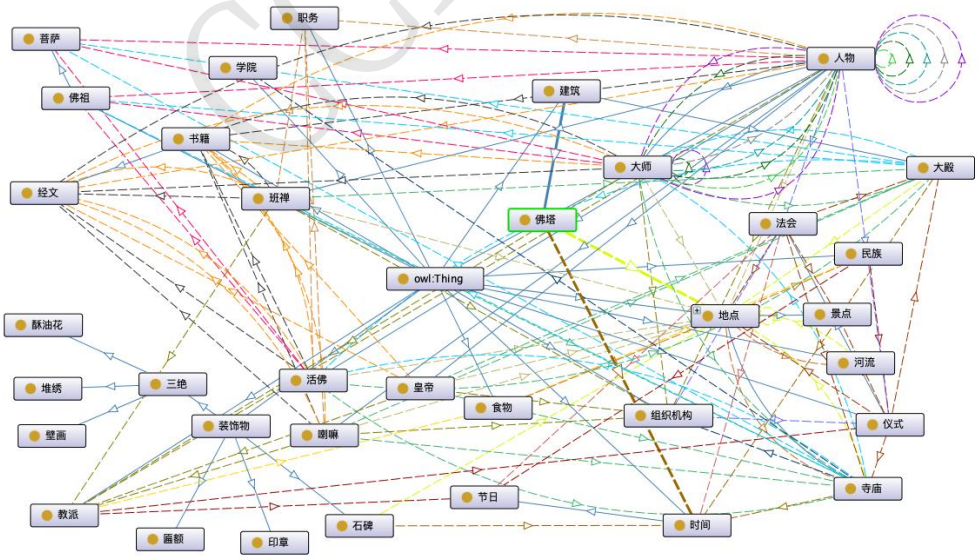


图 2: 塔尔寺图谱本体结构展示

### 3.3 数据标注

数据标注是知识图谱构建过程中最耗费人力与时间的环节，标注成员需具备一定的领域知识且对原始语料进行充足的了解，除此之外还要考虑不同成员的标注习惯以及词语所具有的多样性与歧义性的特性可能造成的标注差异。为此，我们提出了以下建议：在开始标注之前预先制定出一套标注规范，每一位成员充分学习标注规范并了解原始数据，利用标注平台提高标注效率。现有的标注平台种类繁多，不同平台所针对的细化任务也不尽相同，本团队选取了支持多人协同标注的LabelStudio标注平台。标注任务的前期工作包含：（1）对数据进行文本对齐以及章节划分等预标注操作；（2）将原始语料上传至标注平台并拆分为多个部分并分配给每一位标注成员；（3）在平台中创建本体构建任务所设计的31类实体和20种关系；（4）制定标注规范，对标注成员进行培训，提高标注团队结果的一致性。

图 3: LabelStudio标注示例

本团队在命名实体标注过程中制定了以下三点规范：（1）仅可标注已预先定义的31类实体；（2）同一字符串不能重叠标注；（3）除书籍实体的书名号外，尽量不标注其他符号。在关系标注过程中遵循以下两点规范：（1）仅可标注已预先定义的20类关系；（2）优先标注同一句子内的关系，若同一句子内没有关系，允许跨句标注。经过多轮迭代，最终标注成员在塔尔寺原始语料中累积标注命名实体10380个，实体关系3370条。

| 实体1 | 关系类型 | 实体2 |
|---|---|---|
| 人物 | 任职于 | 地点 |
| | | 学院 |
| | | 组织机构 |
| | | 寺庙 |
| | | 职务 |
| | 创建 | 组织机构 |
| | | 学院 |
| | | 仪式 |
| | | 教派 |
| | 出生于 | 地点 |
| | | 时间 |
| | 编著 | 经文 |
| | 供奉 | 佛祖 |
| | 亲属 | 人物 |

| 实体1 | 关系类型 | 实体2 |
|---|---|---|
| 地点 | 位于 | 地点 |
| | 属于 | 地点 |
| | 临近 | 地点 |
| | 拥有 | 寺庙 |
| | | 组织机构 |
| | | 民族 |
| | 举行 | 仪式 |
| 寺庙 | 位于 | 地点 |
| | 供奉 | 佛祖 |
| | 设立 | 组织机构 |
| | | 学院 |
| | 拥有 | 大殿 |
| | | 佛塔 |
| | | 石碑 |

表 1: 部分实体与关系展示

### 3.4 知识抽取

知识抽取是构建知识图谱的核心环节也是难点，本图谱的抽取工作采用了一种基于文本生成的联合抽取模型，在此期间面临着以下问题：（1）数据集规模较小;（2）类别分布高度不平衡，例如“人物”与“寺庙”实体类型的占比远高于“节日”和“职务”类型;（3）三元组重叠如“塔尔寺位于青海省省会西宁市”，其中青海省和西宁市都是塔尔寺所在地。鉴于这些问题，本工作利用(Sun et al., 2019)Erine预训练模型作为编码层结合(Wei et al., 2020)指针网络进行知识联合抽取(Lu et al., 2022)从而解决了类别分布不匹配以及三元组重叠问题，并通过引入提示学习（Prompt learning）(Liu et al., 2021)策略缓解数据标注量较少的问题。

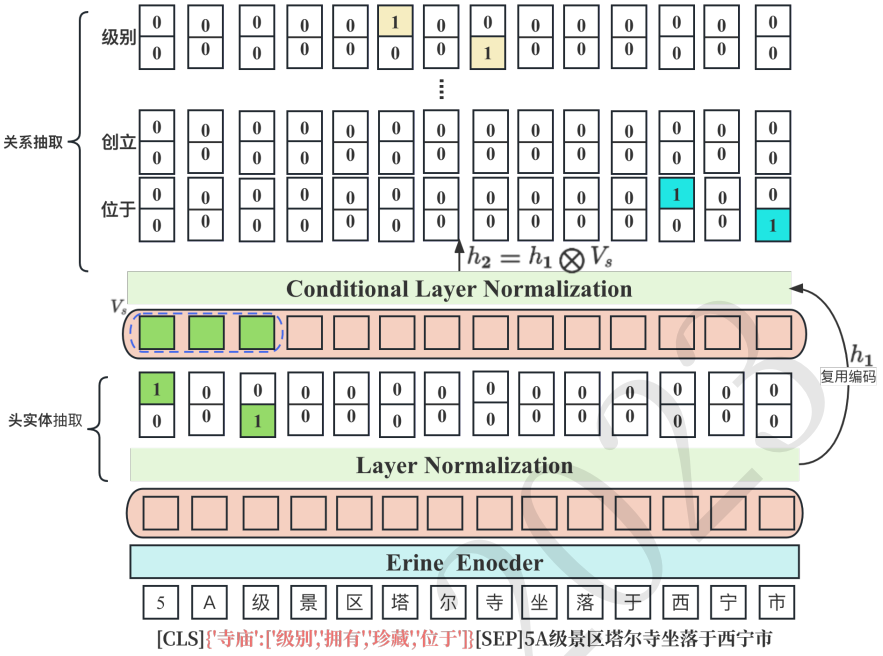


图 4: 基于提示学习与指针网络的联合抽取模型

知识抽取的流程如图4所示，首先依据需要抽取的对象构建模版，模版与文本拼接作为Encoder的输入从而获取输入语句的每一个字符的特征向量表示。之后对实体识别与关系抽取进行建模利用句子的编码信息，抽取出三元组头实体(subject)，具体实现是通过对每个token，使用两个独立的二分类器预测其是否为实体的开始（$p^{s\_start}$）或实体的结束($p^{s\_end}$)，并设定一个阈值，若大于阈值则标记该token为1，否则标记为0。计算如公式（1），公式（2）所示。

$$p^{s\_start} = \sigma(\omega^{s\_start} h_1^i + b^{s\_start}) \tag{1}$$

$$p^{s\_end} = \sigma(\omega^{s\_end} h_1^i + b^{s\_end}) \tag{2}$$

其中，$\omega^{(*)}$可训练权重，$b^{(*)}$ 表示可训练偏置，$\sigma$ 表示sigmoid激活函数，$h_1$为句子编码。主体抽取后进行关系与客体的联合抽取，首先subject的编码$V_s$与句子编码$h_1$进行特征结合得到新的句子编码$h_2$，然后利用新的句子编码$h_2$计算当前subject对应于每一种关系下object的起始token。其抽取过程采用的计算公式与subject抽取计算过程相同，不再赘述。

| | 输入 | 输出 |
|---|---|---|
| 1 | [CLS]寺庙[SEP]5A级景区[MASK]坐落于西宁市 | 塔尔寺 |
| 2 | [CLS]地点[SEP]5A级景区塔尔寺坐落于[MASK] | 西宁市 |
| 3 | [CLS]{’寺庙’:[’级别’,’拥有’,’珍藏’,’位于’]}[SEP]5A级景区[MASK]坐落于西宁市 | （塔尔寺，位于，西宁市）<br>（塔尔寺，级别，5A级） |

表 2: 模版示例

MLM（masked language model）是将一句话中的某个字符用[MASK]替换掉，而后用模型预测这句话中的每一个词，最后利用模型预测被[MASK]替换的词实际上是什么。提示学习（Prompt learning）是伴随输入，给予模型的一个提示模版，用以指导模型接下来应当要做什么任务。也就是说，Prompt learning能够将下游任务改造成预训练模型期望的样子，从而提升模型的表现。在进行知识抽取任务之前，需根据要抽取的对象，制定提示模版，如表2所示，提示模版嵌入在输入语句的头部形成“头模版”，“头模版”可充分激发语言模型的文本生成能力，提升模型的抽取表现。

模版与文本拼接组成Erine（Enhanced Representation through knowledge Integration）模块的输入。其结构与Bert相似，由Embedding层以Transformer层组成，如图6所示。Erine采用了一种如表3所示改进后的MLM策略，将原本以字为单位的Mask方法变为对整个汉语单词Mask，即屏蔽整个词语而不是屏蔽汉字。相较于Bert学习原始语言信号，Erine直接对先验语义知识单元进行建模，增强了模型语义表示能力。

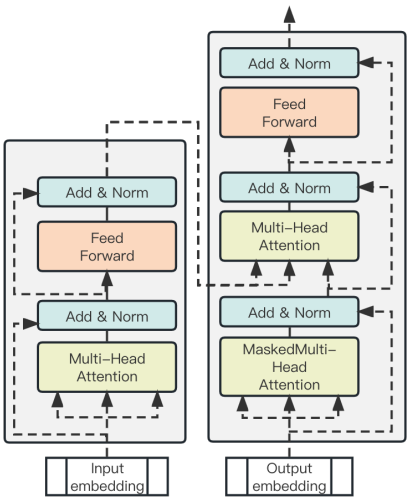


图 5: Transformer模型结构

| 说明 | 样例 |
|---|---|
| 语句 | 5A级景区塔尔寺坐落于西宁市 |
| 原MASK策略 | 5[Mask]级景区塔尔[Mask]坐落于西[Mask]市 |
| WWM策略 | [Mask][Mask][Mask]景区[Mask][Mask][Mask]坐落于西宁市 |

表 3: 全词掩码示例

实体关系抽取的标注策略包含序列标注法和指针标注法，序列标注法的天然的缺陷在于，当实体与上下文中的多个实体存在关系时，只将关系分配给最近的实体(马建红et al., 2021)，无法解决解决三元组重叠的问题。如图5所示，序列标注只能识别出塔尔寺位等级为5A级，无法继续识别出塔尔寺位于西宁市。

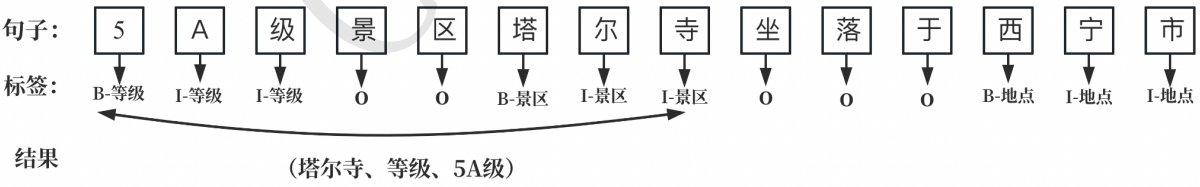


图 6: 序列标注法

本模型的指针网络中采用“0”、“1”对句子中的实体进行标注，具体如图4所示，实体的开始token和结束token标注为“1”，剩余token 标注为“0”。通过建立级联结构，可重复标注塔尔寺对应token，以此有效的解决了三元组重叠问题。

### 3.5 知识融合

由于本图谱原始语料来自于不同著作，不同著作中对同一个客观实体的表述存在差异，对于获取到的知识，需要进行融合加以关联。本图谱的知识融合任务主要是将不同著作中对同一实体的不同语义表述，关联到同一客观实体上，从而实现同一实体的多名语义消歧。具体的我

们首先统计出各个实体的词频，利用Bert学习实体的表示，融合word2vec的表示和编辑距离(梅筱and 刘海鹏, 2010)一起来计算两个实体的相似度。如果两实体的相似度大于设定的阈值，则将词频较低的实体融合近词频较高的实体中。

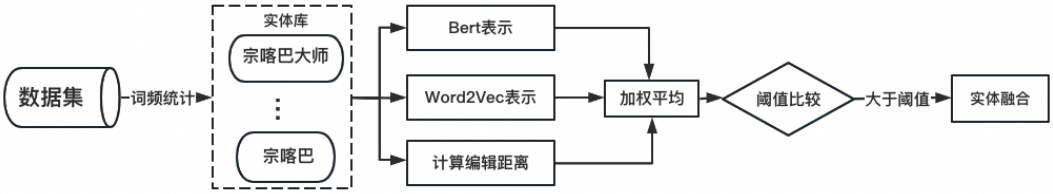


图 7: 知识融合流程

如图7所示，首先统计出实体"宗喀巴大师"与"实体宗喀巴"的频率，经过对词语的Bert表示，以及word2vec的表示和编辑距离进行加权平均，得出低频词"宗喀巴大师"与高频词"宗喀巴"的相似度大于阈值，则认为其二者表述的是客观世界的同一实体，最后将"宗喀巴大师"实体融合到词频更高的"宗喀巴"实体中。

## 4 塔尔寺知识图谱构建

### 4.1 实验

本实验采用Human-in-loop方式(Wang et al., 2021)，即数据标注与模型训练循环交替进行，每一轮标注后对模型进行一次训练，一次标注加上一次模型训练构成一轮迭代，凭借这种方式可清晰的观察出数据集标注量对实验效果的影响，并且可以依据上一轮次的实验结果，开展新一轮次更具针对性的标注工作。

本文采用精确率（P），召回率（R）以及F1-score（F1）值作为模型结果的评判指标。具体计算公式如公式（3）、公式（4）、公式（5）所示，其中，TP表示模型能正确检测出的实体个数、FP表示模型检测到的无关实体个数、FN表示模型未检测到的实体的个数。

$$P = \frac{TP}{TP + FP} \times 100\% \tag{3}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{4}$$

$$F1 = \frac{2PR}{P + R} \times 100\% \tag{5}$$

本实验使用Tesla PG503-216显卡对模型进行训练，显卡内存128G，服务器系统Linux ubuntu 5.4.0-125-generic，python版本3.7.16，训练框架paddlepaddle-gpu 2.4.1。

经过5轮迭代后。标注成员最终在塔尔寺原始数据集中共标注命名实体10380个，实体关系3370条。各实验周期的结果如表4所示。

| 轮次 | 实体标注量 | 关系标注量 | 实体抽取结果 | | | 关系抽取结果 | | |
|---|---|---|---|---|---|---|---|---|
| | | | P | R | F1 | P | R | F1 |
| 迭代-1 | 4062 | 835 | 55% | 62.8% | 58.6% | 52.2% | 64.7% | 57.8% |
| 迭代-2 | 6015 | 1346 | 59.3% | 61.2% | 62.5% | 60.3% | 66.6% | 61.5% |
| 迭代-3 | 8093 | 1677 | 64.2% | 60% | 62.1% | 68.4% | 68.4% | 68.4% |
| 迭代-4 | 9052 | 2033 | 77.7% | 87% | 82.3% | 64.9% | 85.7% | 80% |
| 迭代-5 | 10380 | 3370 | 86.7% | 86.1% | 86.4% | 80.6% | 83.3% | 81.9% |

表 4: 各迭代阶段实验结果

实验结果表明，伴随数据标注量的逐渐增加，实验的结果也在稳步提升。最终实体抽取任务与关系抽取任务在各项评价指标下的表现都超过了80%，不同本体类型与不同关系类型之间

的抽取表现相差较大，其抽取表现与权重成正相关关系，本图谱中核心类型实体与关系的实验结果如表5所示。

| 实体 | P | R | F1 | weight | 关系 | P | R | F1 | weight |
|---|---|---|---|---|---|---|---|---|---|
| 学院 | 95.2% | 76.9% | 85.1% | 3.7% | 珍藏 | 100% | 67% | 43% | 7.4% |
| 民族 | 93.3% | 56% | 70% | 2.1% | 编著 | 100% | 50% | 67% | 4.3% |
| 寺庙 | 86.7% | 83.3% | 84.8% | 11.7% | 位于 | 94.5% | 62.2% | 66.7% | 11.1% |
| 法会 | 85.7% | 66.7% | 75% | 8.9% | 供奉 | 89.4% | 56.2% | 70% | 6.2% |
| 教派 | 84.6% | 78.5% | 81.4% | 2.9% | 临近 | 83.6% | 79.7% | 81.2% | 19% |
| 大师 | 83.3% | 71.4% | 76.9% | 7.2% | 称呼 | 82.9% | 87.1% | 85% | 5.5% |
| 时间 | 82.3% | 80.1% | 81.2% | 5.7% | 拥有 | 80.1% | 66.7% | 72.7% | 6.6% |
| 佛像 | 81.4% | 84.6% | 83.1% | 9.7% | 属于 | 77.7% | 77.7% | 77.7% | 7.3% |
| 书籍 | 80.6% | 83.3% | 82% | 6.6% | 生于 | 75% | 51.7% | 61.2% | 2.2% |
| 人物 | 77.8% | 87.5% | 82.3% | 14.6% | 亲属 | 70.4% | 51.1% | 59.3% | 3.1% |

表 5: 实体关系抽取结果

## 4.2 对比实验

为验证实验的有效性，本实验分别与流水式抽取以及联合抽取相比较。流水式抽取中采用Bert结合Bilstm以及条件随机场(CRF)作为实体抽取任务的模型，并采用Bert和MLP作为关系抽取任务的模型。联合式抽取采用先利用Global Pointer抽取subject的首尾(i, j)和object的首尾位置(i, j)，然后利用Global Pointer抽取每一种关系p的实体的头部所对应的位置(hi, hj)和实体的尾部tail所对应的位置(ti, tj)组合，最终输出交集的结果。从表6可以看出，联合抽取的表现均优于流水式抽取，GPlinker相对流水式抽取的表现有了明显提升。而本模型使得实体抽取以及关系抽取的表现进一步的提升，图8展示了对比实验在关系抽取任务中的表现。

| 方式 | 模型 | 任务类型 | F1 |
|---|---|---|---|
| Pipline | Bert+Bilstm+CRF | 实体抽取 | 70.5% |
| | Bert+MLP | 关系抽取 | 69.5% |
| Joint | GPlinker | 实体抽取 | 84.9% |
| | | 关系抽取 | 77.7% |
| Joint | 本文模型 | 实体抽取 | 86.4% |
| | | 关系抽取 | 81.9% |

表 6: 对比实验结果

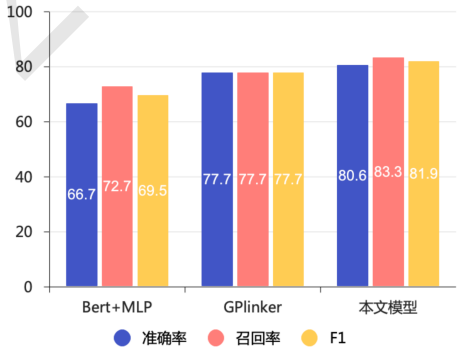


图 8: 对比实验关系抽取结果

## 4.3 错误分析

经过对实验结果和实验语料进的分析，可得出错误原因主要由以下两方面造成。首先是关系分布的不平衡，塔尔寺知识图谱所定义的20类关系子类型中，‘位于’和‘临近’等9种常见关系的语料数量占比达到了百分之70，其余11类关系对应语料的总和占比仅百分之30，这导致了大部分的关系和实体对的可用样例仍然较少，而神经网络模型作为典型的需要大量训练数据支撑的技术，在训练样例过少的情况下各项评价指标都受到极大影响。其次塔尔寺知识图谱原始语料的非结构化信息比较复杂，著作中包含较多的代词和长难句，这些对信息抽取都造成了负面影响。实验结果中有部分关系类型例如’珍藏’与‘编著’虽占比较少，但也取得了较好的抽取表现，这是由于该类实体在语料中集中分布于特定章节。

## 4.4 知识存储

本图谱利用Neo4j图数据库作为知识存储工具，实现了知识查询与更新以及界面化展示等功

能(冯俐, 2019)，可为图谱的下游互联网任务提供支持。塔尔寺部分知识在数据库中的存储形式如图8所示，经过人工校对，剔除掉6709条错误信息，最终本图谱针对塔尔寺31类实体以及20种关系所构建的知识图谱共包含4705个节点及17386条关系。

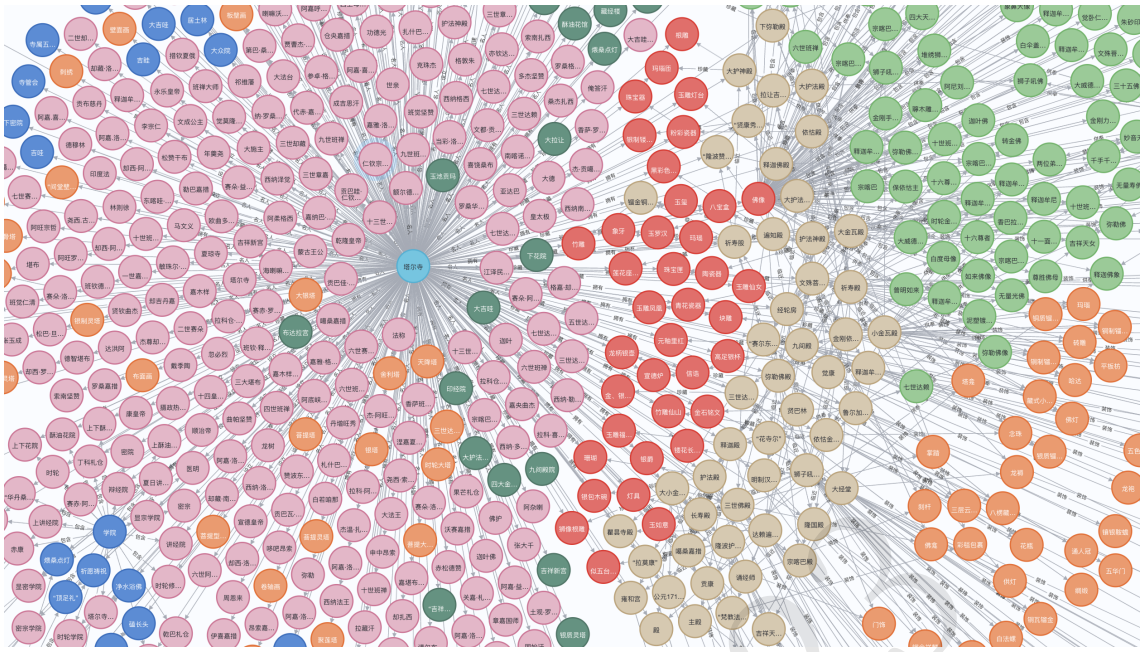

图 9: 塔尔寺知识图谱存储展示

### 4.5 图谱应用

伴随文旅产业的复苏，青海省的景区也再次进发出活力，2023年五一假期期间，塔尔寺景区游客量已经达到2019年同期的118.6%，面对激增的游客量，景区服务水平亟待提升。我们将通过引入更多的著作构建出足以支撑塔尔寺问答助手的知识图谱，该问答助手除了为游客介绍各个景点的风貌特征外还可以讲解塔尔寺的历史典故，并且该问答助手可以作为教育应用，帮助游客提前了解塔尔寺文化，推广塔尔寺丰厚的文化传播，实现塔尔寺文化的传承保护。

## 5 总结

本文的意义在于：1）立足数字人文，提出一套针对青藏文化保护工作的范例，为日后青藏文化保护工作的开展提供了新思路。2）将提示学习与基于文本生成式的实体关系联合抽取模型结和，实现了在低资源情况下对塔尔寺相关知识的抽取。3）完整的介绍了塔尔寺知识图谱的构建流程，弥补了人文领域有塔尔寺的结构化数据不足的问题，满足了互联网任务需求，实现了对传统文化的传承保护。

## 参考文献

M. Abdellatif, M. S. Farhan, and N. S. Shehata. 2017. Overcoming business process reengineering obstacles using ontology-based knowledge map methodology. *Future Computing Informatics Journal*, page S2314728817300296.

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Y. Lu, Q. Liu, D. Dai, X. Xiao, H. Lin, X. Han, L. Sun, and H. Wu. 2022. Unified structure generation for universal information extraction.

M. Miwa and M. Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures.

Paulheim and Heiko. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic Web*.

Y. Sun, S. Wang, Y. Li, S. Feng, and H. Wu. 2019. Ernie: Enhanced representation through knowledge integration.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. *arXiv preprint arXiv:2010.13415*.

Z. J. Wang, D. Choi, S. Xu, and D. Yang. 2021. Putting humans in the natural language processing loop: A survey.

Z. Wei, J. Su, Y. Wang, Y. Tian, and Y. Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

S. Zheng, F. Wang, H. Bao, Y. Hao, P. Zhou, and B. Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme.

傅柱, 王曰芬, and 孙铭丽. 2013. 本体存储技术研究综述. 情报理论与实践, 36(9):6.

冯俐. 2019. 基于neo4j图数据库构建中学语文诗词知识图谱. 陕西师范大学.

刘峰贵, 王锋, 张海峰, 周强, 陈琼, and 李春花. 2012. 青藏高原文化旅游资源开发探讨. 青海社会科学, (5):6.

刘泽权. 2021. 数字人文视域下名著重译多维评价模型构建. 中国翻译.

刘济源. 2019. 旅游领域知识图谱的构建及应用研究. Ph.D. thesis, 浙江大学.

周莉娜, 洪亮, and 高子阳. 2019. 唐诗知识图谱的构建及其智能知识服务设计. 图书情报工作, 63(2):10.

孙乃荣, 雷芳, 侯晓丹, and 陈杭. 2022. 数字人文视域下河北省文化和旅游外宣网站高质量发展对策研究. 西部旅游, (16):77–79, 8.

张卫, 王昊, 邓三鸿, and 张宝隆. 2021. 面向数字人文的古诗文本情感术语抽取与应用研究. 中国图书馆学报, 47(4):19.

张宇飞, 李腾, and 贾东立. 2022. 河北省旅游景点知识图谱的构建. 计算机与数字工程, (007):050.

梅筱and 刘海鹏. 2010. 基于编辑距离结合词性的词相似度算法.

翁冉, 何世群, 杨秀璋, and 罗子江. 2023. 国内数字人文领域热点及趋势探析. 河南图书馆学刊, 43(1):3.

聂欣晗and 张亮玉. 2021. 论四大名著的文化旅游开发. 旅游纵览, No.354(148-152).

马建红, 魏字默, and 陈亚萌. 2021. 基于信息融合标注的实体及关系联合抽取方法. 计算机应用与软件, 38(7):8.