

基于知识图谱支持的西藏交通问答系统研究

史雅琴¹ 樊越¹ 胡心龙¹ 次仁欧珠¹ 程刚^{1*}
SHI Yaqin FAN Yue HU Xinlong Cirenouzhu CHENG Gang

摘要

针对西藏地区地形复杂、交通信息分散、传统信息获取方式效率低下的问题,文章构建了一个基于知识图谱的西藏交通问答系统。该系统利用知识图谱技术整合西藏地区交通相关的多源异构数据,构建结构化的交通知识库;并结合自然语言处理技术,实现对用户自然语言查询的智能问答功能。该系统能够为游客提供精准、实时的交通信息,有效解决传统信息渠道存在的信息分散、更新滞后等问题,对于提升西藏地区旅游信息服务的质量与效率具有重要的现实价值。

关键词

西藏交通;知识图谱;问答系统;自然语言处理

doi: 10.3969/j.issn.1672-9528.2026.02.044

0 引言

西藏地区以其丰富的旅游资源和独特的文化魅力吸引着众多游客,然而复杂的地形条件和相对滞后的交通基础设施建设给游客出行带来了显著挑战^[1]。目前游客获取交通信息主要依赖传统搜索引擎和旅游网站,这些渠道存在信息分散、更新滞后且准确性难以保证等问题,严重影响了旅游体验。在该背景下,开发一个能够整合多源数据、提供精准实时交通信息的智能问答系统具有重要的实践价值。知识图谱技术为这一需求提供了理想解决方案,该技术通过结构化方式组织现实世界中的实体、属性及其相互关系,构建起庞大的语义网络^[2]。基于知识图谱的问答系统能够将自然语言查询转化为结构化检索,从语义网络中精准提取相关实体、属性和关联信息,最终生成准确、简洁的应答,这种智能化服务模式将有效提升西藏地区旅游信息服务的质量和效率^[3]。

1 知识图谱的构建

构建知识图谱的数据主要分为结构化数据、半结构化数据和非结构化数据。其中半结构化数据主要来源于网站上通过爬虫爬取的按关键字分类好的数据。本研究采用的半结构化数据主要来自西藏交通部门、旅游部门发布的官方交通路

线、地点信息,以及旅游网站、地图服务和点评平台提供的商户信息与用户评价等数据。通过数据爬取和清洗流程,我们剔除了重复、错误和不完整的数据,并将其转换为标准化的CSV格式,既作为图谱数据,同时对数据进行标注,进行实体识别和关系抽取。

知识抽取是从非结构化数据中提取结构化知识的过程,主要分为流水线抽取和联合抽取,流水线抽取先对文本进行实体识别,抽取出头实体和尾实体,再将两个实体和文本放入另一个模型中进行关系抽取,即文本分类对应的关系^[4]。由于实体识别和关系识别采用了不同的模型,因此灵活性较差,存在错误传播和上下文信息丢失的问题,例如实体识别错误会导致关系抽取失败,并且每个模块只处理局部信息,缺乏全局上下文理解,模型会忽略长距离依赖和跨句信息。联合抽取是一种同时进行实体识别和关系抽取的方法,旨在通过共享信息和联合建模,克服传统流水线方法的弊端,模型通过共享输入文本的表示,使用一个损失函数,同时处理两个任务,以此来提升知识抽取的性能。

本文使用 prgc 模型来进行关系抽取,其核心思想是通过潜在关系预测和全局对应机制,减少冗余计算并提升模型对复杂语境的理解能力^[5]。主要分为三个部分,首先预测文本中可能存在的关系,通过构建一个全局指针对所有可能存在关系的主宾位置进行标记;其次捕捉实体和关系之间的全局依赖,如果句子中存在某一特定关系,那么就将这个关系所在的位置进行标记;最后在潜在关系和全局对应的基础上,进行实体识别和关系分类,针对存在的多个关系,计算出此关系的头尾实体位置,将其作为最终的三元组数据进行输出。模型整体如图1所示。

1. 西藏大学工学院 西藏拉萨 854000

[基金项目] 大学生创新创业训练项目“基于多源地理信息的智能交通问答系统设计与实现基金”(2025XCX022);西藏自治区科技计划揭榜挂帅项目(XZ202303ZY0005G);中央引导地方科技发展资金项目“拉萨城市交通精细化治理研究”(LSKJ202461)

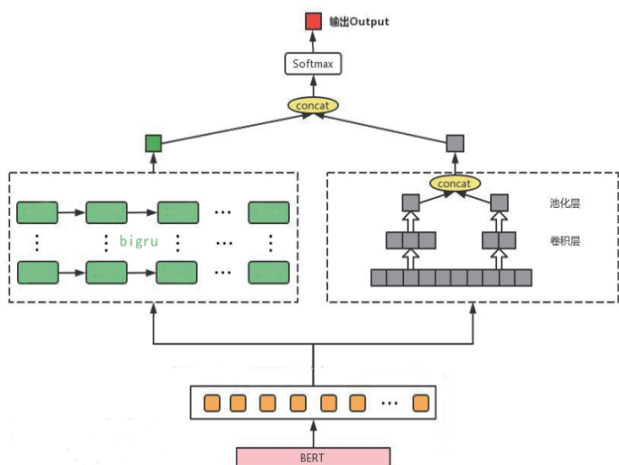


图4 融合 TextCNN 与 BiGRU 的特征融合模型

TextCNN 是一种用于文本分类的卷积神经网络模型，其核心思想是将文本数据视为一维的序列，通过使用多个不同大小的卷积核，对输入矩阵进行卷积操作。每个卷积核可以看作是一个特征提取器，用于捕捉文本中的局部特征，接着对每个卷积核的输出进行最大池化操作，提取每个特征图中的最显著特征。最大池化操作可以减少特征的维度，同时保留最重要的信息，将池化层的输出拼接成一个向量，然后通过一个或多个全连接层进行分类。

GRU 是一种改进的循环神经网络结构，通过引入门控机制来控制信息的流动^[12]。但其结构更加简洁，只有两个门，更新门和重置门，其中更新门决定当前时刻的状态有多少来自前一时刻的状态，有多少来自当前时刻的输入。重置门决定前一时刻的状态对当前时刻的影响程度。

3 实验结果分析

3.1 实验参数设置

本文采用的模型通过 TENSORFLOW 平台进行搭建，实验环境如表 1 所示。

表 1 实验环境

配置	参数
操作系统	Windows 10
CPU	AMD Ryzen 7 6800H with Radeon Graphics 2.90 GHz
GPU	NVIDIA GeForce RTX 4060
编程语言	Python 3.7
内存	32 GB

本文将网上爬取的有关西藏交通的信息进行整合，通过标注精灵对数据进行标注。在知识抽取实验中，选用 BERT 作为编码层，输入维度为 768，学习率设置为 $1e-4$ ，共训练 30 个 epoch。在命名实体识别实验中，将 BiLSTM 隐藏层维度设置为 128，批量大小设置为 16，学习率设置为 $2e-5$ ，训练 10 个 epoch。在文本分类实验中，CNN 使用的卷积核大小为 [2, 3, 4]，每种卷积核数量为 100，GRU 的隐藏层维度为 128，Dropout 为 0.2，共训练 10 个 epoch。

3.2 实验结果分析

本文使用精确率、召回率、 F_1 值对实验结果进行评估，其中精确率衡量模型预测为正类的样本中有多少是真正的正类，召回率衡量实际为正类的样本中有多少被模型正确预测为正类， F_1 值是精确率和召回率的调和平均值，用于综合评估模型的性能。其计算公式为：

$$P = \frac{TP}{TP + FP} \quad (8)$$

$$R = \frac{TP}{TP + FN} \quad (9)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \quad (10)$$

本文使用的知识抽取模型通过与 TPLinker 和 CasRel 模型进行对比，实验结果如表 2 所示。

表 2 知识抽取模型

模型	P/%	R/%	F_1 /%
CasRel	75.73	73.91	74.82
TPLinker	77.61	76.47	77.02
PRGC	78.89	77.43	78.14

由表 2 可以看出，本文模型比 CasRel 和 TPLinker 模型 F_1 值分别提高了 3.32% 和 1.12%，其原因是 CasRel 采用级联二元标记框架，需要对每个关系类型进行独立的实体识别和关系分类，导致计算冗余，TPLinker 通过标记对链接的方式进行实体和关系的联合抽取，需要对所有可能的标记对进行处理，计算复杂度较高，而 PRGC 通过潜在关系预测和全局对应机制，减少了冗余计算，捕捉了全局依赖，从而提升了模型的效率和性能。

本文使用的命名实体识别模型通过与 BERT_SOFTMAX、BiLSTM_CRF、BERT_CRF 进行对比，实验结果如表 3 所示。

表 3 实体识别模型

模型	P/%	R/%	F_1 /%
BERT_SOFTMAX	87.92	88.39	88.14
BiLSTM_CRF	90.28	91.34	90.80
BERT_CRF	90.83	91.67	91.24
BERTA_BiLSTM_CRF	91.37	92.06	91.70

由表 3 可以看出，加入 CRF 后的 BERT 模型比不加 CRF 的 F_1 值提高了 3.1%，主要原因在于 CRF 的引入能够显式地建模标签之间的依赖关系，从而解决 BERT 在序列标注任务中的一些局限性；BERT_BiLSTM_CRF 比 BiLSTM_CRF、BERT_CRF 模型的 F_1 值分别高 0.9% 和 0.46%，说明通过引入 BiLSTM，可以增强序列依赖关系的建模能力，而引入 BERT 可以提升了上下文表示能力和泛化能力。

本文使用的文本分类模型通过与 BERT_TextCNN 和 BERT_BiGRU 进行对比，实验结果如表 4 所示。

表4 文本分类模型

模型	P/%	R/%	F_1 /%
BERT_BiGRU	83.19	82.64	82.92
BERT_TextCNN	85.76	84.91	85.34
本文模型	86.87	85.34	86.08

由表4可以看出,进行特征融合后的文本分类模型比只使用卷积神经网络和门控循环网络模型的 F_1 值分别提高0.74%和3.16%,其原因在于TextCNN擅长捕捉局部特征,能够提取文本中的关键短语或模式,而BiGRU擅长捕捉序列中的长期依赖关系,能够建模文本的上下文信息,通过将二者的特征进行融合,可以结合二者的优势,弥补各自的不足,从而提升模型的性能。

4 系统设计与实现

将命名实体识别得到的实体数据与文本分类得到的意图数据放入cypher语句中得到最后的查询语句,例如match(n:地点) where n.address = 万达广场 return n.人流密度表明用户想知道实体万达广场的人流密度信息,即用户的意图。问答系统前端采用Django框架,界面由html,css编写,前后端交互通过JavaScript实现,当用户的问句出现在输入框中时,通过发送按钮,系统将会把问答结果返回给用户界面,如图5所示。



图5 问答系统界面

5 结论

本项目成功构建了一个基于知识图谱的西藏交通问答系统,实现了交通信息的智能化获取。该系统利用知识图谱技术整合多源异构数据,构建了结构化的西藏交通知识库,并结合自然语言处理技术实现了智能问答功能,为用户提供准确、及时、全面的交通信息。未来会继续完善知识图谱:不断扩充和更新西藏交通知识图谱,提高知识库的覆盖范围和准确性。同时优化问答系统:提升问答系统的理解和推理能力,支持更复杂的自然语言问句。

参考文献:

- [1] 刁成林.当代中国西藏交通研究综述[J].西藏研究,2022(2):148-160.
- [2] 张西硕,柳林,王海龙,等.知识图谱中实体关系抽取方法研究[J].计算机科学与探索,2024,18(3):574-596.
- [3] 王志明,郑凯.基于BERT的中文医疗问答系统[J].计算机系统应用,2023,32(6):115-120.
- [4] 张仰森,刘帅康,刘洋,等.基于深度学习的实体关系联合抽取研究综述[J].电子学报,2023,51(4):1093-1116.
- [5] ZHENG H Y, WEN R, CHEN X, et al. PRGC: potential relation and global correspondence based joint relational triple extraction[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Brussels: ACL, 2021:6225-6235.
- [6] 文辉,于敬,周明星.基于知识图谱的失效分析系统设计与应用[J].电子技术与软件工程,2022(16):243-248.
- [7] 权义琦,张海涛,杨斌,等.基于深度学习的命名实体识别综述[J/OL].微电子学与计算机,1-11[2026-01-28].https://link.cnki.net/urlid/61.1123.tn.20250312.0854.002.
- [8] 王鑫澳,陈珂,寿黎但,等.基于联邦学习的BERT模型高效训练框架[J].软件学报,2025,36(9):4110-4133.
- [9] 姚松林,梁敦毫,罗振营.基于BERT-BiLSTM-CRF模型的荔枝命名实体识别[J].中国科技信息,2025(3):110-113.
- [10] 高珊,李世杰,蔡志平.基于深度学习的中文文本分类综述[J].计算机工程与科学,2024,46(4):684-692.
- [11] 苏易礪,李卫军,李贯峰,等.基于门控特征融合的双通道文本分类模型[J/OL].计算机工程与应用,1-15[2026-01-28].https://link.cnki.net/urlid/11.2127.tp.20250225.1637.018.
- [12] 陈农田,李俊辉,满永政.基于改进CNN-BiGRU-att模型的文本分类研究[J].昆明理工大学学报(自然科学版),2022,47(1):30-37.

【作者简介】

史雅琴(2004—),女,青海海东人,本科在读,研究方向:交通运输。

樊越(2004—),男,山西忻州人,本科在读,研究方向:交通运输。

胡心龙(2003—),男,河南周口人,本科在读,研究方向:自然语言处理。

次仁欧珠(1970—),男,西藏日喀则人,本科,副教授,研究方向:交通运输工程。

程刚(1981—),通信作者(email:176064239@qq.com),男,四川泸州人,博士,副院长、教授、博士生导师,研究方向:交通运输。

(收稿日期:2025-08-13 修回日期:2026-02-03)