基于深度学习的道路交通风险因素知识图谱构建研究

张丽岩[1] 顾欣怡[1] 马 健[1]
ZHANG Liyan GU Xinyi MA Jian

摘 要 深度学习技术的进步带来了知识图谱新的发展，使知识图谱能够在智能搜索、智能问答、个性化推荐等多个领域得到认可。BGA-CRF 模型以道路交通事故数据为基础，通过知识抽取和构建技术等步骤，提取出关键信息并确保数据的一致性，得到影响道路交通事故的主要风险因素 11 种，借助 Neo4j 图数据库技术构建融合道路交通事故数据风险因素的知识图谱，从而实现数据的存储与可视化，并开发问答系统以便于更好地理解和分析相关数据，提升检索信息的效率，从而扩展了知识图谱的应用范围。

关键词 知识图谱；问答系统；交通事故；深度学习；Neo4j；可视化

doi：10.3969/j.issn.1672-9528.2025.02.034

0 引言

近年来，全国汽车保有量持续增长。据统计，截至 2023 年底，全国的民用汽车保有量已增至 3.36 亿辆，相较于 2022 年增加 1714 万辆 [1]。随着汽车保有量的不断增加，道路交通事故的发生率也呈现上升趋势。2023 年，全国共记录 25.47 万起道路交通事故，相较于 2022 年增长了 8%。数据表明，道路交通安全领域正面临着极为严峻的挑战。因此，强化交通安全管理举措、切实降低事故发生率，已然成为当下高度关注并亟需采取切实有效措施加以解决的重要课题。 因此，通过构建道路交通事故领域风险因素知识图谱，利用自然语言处理技术来处理事故数据中的非结构化信息，以提取关键信息并将其永久存储，形成事故知识库至关重要。在此基础上设计智能问答系统，不仅能促进对事故信息的深入挖掘，更有助于分析事故的风险因素和成因，为交通安全管理提供更为精确的数据支持和决策依据。

1 逻辑架构

知识图谱嵌入技术在推荐系统的应用中已经受到广泛关注。通过将知识图谱的结构化信息整合到推荐算法中，可以显著提升推荐结果的个性化水平 [2]。

知识图谱架构主要分为数据层与模式层。数据层以三元组的形式表示，负责存储实际的道路交通事故数据，确保知识图谱的灵活性和可扩展性；模式层作为知识图谱的核心所在，定义知识图谱的概念模型和逻辑结构。在交通事故分析领域，将事故实体和其成因关系映射到向量空间中，通过向量间的匹配和运算来捕捉和解析语义信息，从而挖掘出潜在的交通事故信息 [3]。

2 技术架构

在技术层面，构建知识图谱所涉及的关键技术包括实体抽取、关系抽取、知识检索等方面 [4]，共同构成知识图谱的核心技术体系，如图 1 所示。

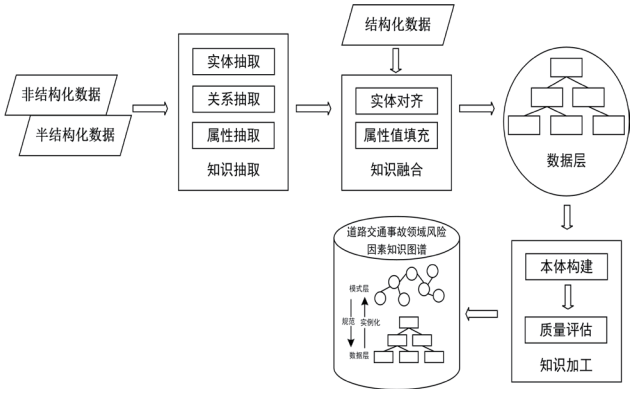

图 1 道路交通事故领域风险因素知识图谱构建

2.1 数据来源及构建思路

本文选取某市 2023 年道路交通事故实例数据作为分析对象，共计 9937 条记录。通过对事故数据的整理和分析，识别出 11 种不同的实体类型和 9 种关系类型。为进一步增强图谱的实用性并拓展其应用范围，构建了一个融合三大商业车险信息的道路交通事故风险因素知识图谱，并开发相应的智

1.1. 苏州科技大学土木工程学院 江苏苏州 215011

[基金项目] 本研究由江苏省研究生实践创新项目（SJCX20_1117、SJCX21_1420、KYCX21_2999）；江苏省建设体系项目（2020ZD14、2018ZD258）；苏州社会科学基金（Y2020LX017、Y2020LX025）；江苏省大学哲学社会科学项目（2018SJA1348、2023SJYB1420）资助

能问答系统。这一系统旨在帮助用户快速获取交通事故风险因素的相关信息并及时预防风险。图 2 展示了该知识图谱的概念模型。

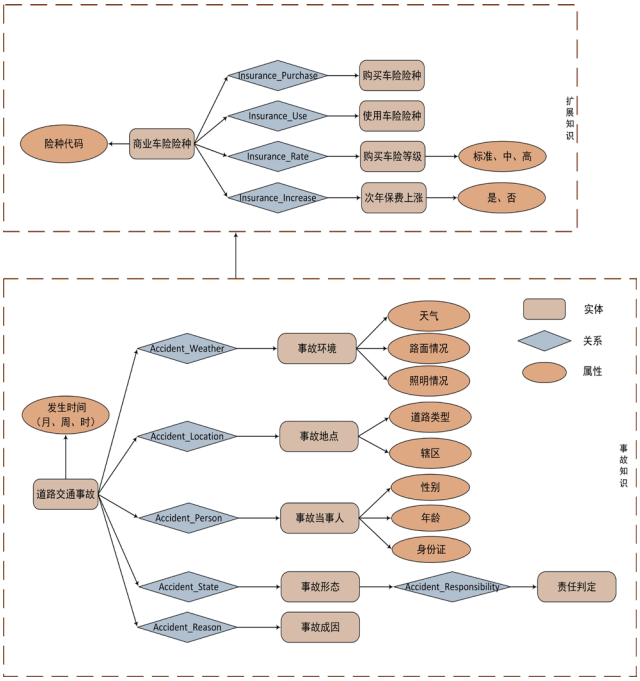

图 2 道路交通事故风险因素知识图谱模型

2.2 数据标注

本文中，首先对文本数据中的实体进行 BIO（begin inside outside）标注。完成标注后，将文件以 .ann 格式导出。随后，利用 Python 脚本读取 .ann 文件中的标注信息，从中提取每个字符的 BIO 标签，从而构建基于 BIO 标注的交通事故文本数据序列。在这一标注体系中，“B”代表实体的起始字符，“I”代表实体的内部字符，“O”则代表非实体字符。数据标注示例如表 1 所示。

表 1 BIO 标注示例

实体类型	起始字符标志	中间字符标志	结尾字符标志
交通事故	B-ACCID	I-ACCID	L-ACCID
事故地点	B-LOC	L-LOC	I-LOC
事故成因	B-REA	I-REA	I-REA
事故形态	B-TYPE	I-TYPE	I-TYPE
非实体类型	O	O	O

2.3 改进后的 BGA-CRF 模型

在对主流的 BERT-BiLSTM-CRF 模型 [5] 进行优化的过程中，对模型的核心组件 BiLSTM 进行升级，将其替换为 BiGRU，并引入注意力机制，以更好地捕捉文本中的关键信息，提高对结构化信息的提取精度。

改进后的 BGA-CRF 模型主要由图 3 四个关键模块构成，模型的工作流程如下：

首先，利用 BIO 标注方法处理文本数据，构建一个专门针对道路交通事故风险因素的语料库。接着，将该语料库中的文本送入 BERT 模块，以生成词向量；这些词向量随后被送入 BiGRU 层进行深入的特征学习。在这一过程中，引入了自注意力机制来增强模型的性能。之后，CRF 模块对特征进行序列解码。最终，模型综合预测并输出最优的标注序列。整体模型结构设计如图 3 所示。

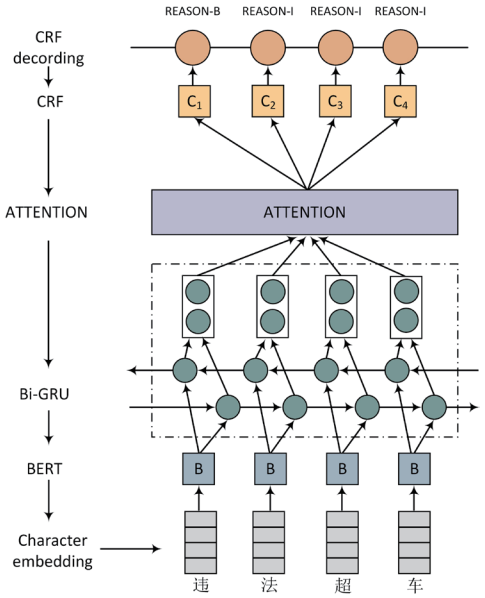

图 3 BGA-CRF 模型架构

2.3.1 BERT 模型

BERT 作为自然语言处理领域内的预训练语言模型，具备学习词汇间以及句子间关系的能力 [6]。对比其他模型，BERT 的优秀之处就在于能探寻文本中的逻辑，面对文本间可能蕴含的真实语境，BERT 模型显现出了优越的特性。模型的结构如图 4 所示。

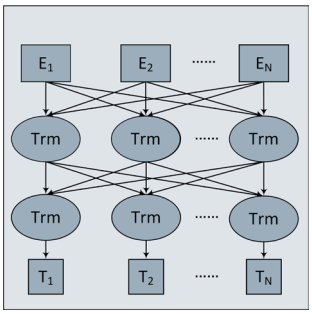

图 4 BERT 模型结构

2.3.2 BiGRU 模型

GRU 内部的两个门：更新门和重置门，负责决定新输入信息与先前记忆的结合方式，以及确定保留多少先前记忆到当前时间步 [7]。这两个门控机制共同决定了哪些信息将作为 GRU 的输出 [8]。双向 GRU（BiGRU）由两个 GRU 层构成，

能够捕捉更多信息，同时处理前向与后向部分的序列有助于扩大模型的视野，更好地理解文本序列结构，对预测能力有一定的提高。模型的结构如图 5 所示。

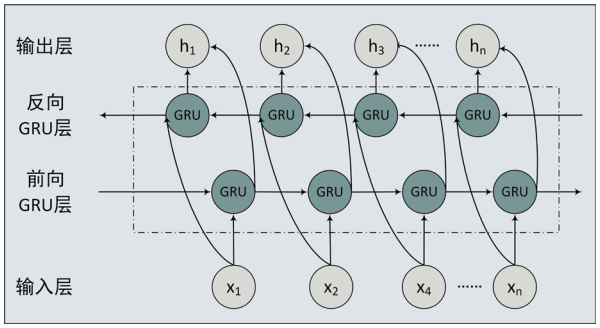

图 5 BiGRU 模型结构

2.3.3 自注意力机制层

自注意力机制也称为内部注意力机制，能够对序列中的不同位置赋予不同的关注程度。每个元素的重要性都由其对应的注意力权重来决定。通过添加自注意力机制模型，可以并行处理序列中的所有元素，去除时间步的影响，捕捉序列中任意两个元素之间的联系，提高处理效率。模型的结构如图 6 所示。

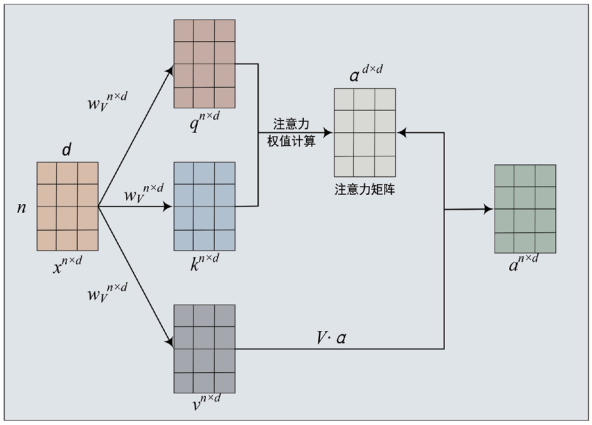

图 6 自注意力机制工作原理

2.3.4 CRF 层

为了增强实体识别的准确性，在 BiLSTM 层之后加入 CRF[9] 层，通过对序列向量施加计算约束，确保输出与预设标签之间的一致性。基于观测到的序列 x，CRF 层计算输出标签序列 $y = [y_1, y_2, \cdots, y_n]$ 的条件概率，其分数函数定义为：

$$f_{ij} = (y_{n-1}, y_n, x_n) = [y_{n-1} = i \text{ and } y_n = j] \quad (1)$$

$$\text{Score}(y,x) = \sum_{n=1}^{N} A_{y_{n-1}y_n} + \sum_{n=1}^{N} \lambda_{y_n} \quad (2)$$

式中：A 为网络层输出的转移分数矩阵；$A_{y_{n-}y_n}$ 是从标签 y_{n-1} 转移到 y_n 的权重；N 是文本长度；λ_{y_n} 是关于单个标签 y_n 的相关权重。CRF 模型计算标签序列的概率分布得分矩阵具体实现为：

$$P(Y|X) = \frac{\exp\left(\text{score}(y,x)\right)}{\sum_{y'} \exp\left(\text{score}(y',x)\right)} \quad (3)$$

2.4 实验与结果分析

在实验阶段，本文对比了 BiLSTM 模型、BERT-BiLSTM-CRF 模型以及 BGA-CRF 模型的性能。为了评估这些模型的效果，采用了精确率 P、召回率 R 和 F_1 分数作为衡量标准。计算公式分别为：

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

$$F_1 = \frac{2PR}{P + R} \quad (6)$$

分析表 2 的数据发现，采用组合模型相较于单一模型能够带来更佳的性能表现，这体现在模型的精确率 P、召回率 R 和 F_1 分数随着模型的优化而稳步提升。在所有比较的模型中，BGA-CRF 模型在 3 个关键评价指标上均展现出最高的数值，分别为 96.93%、96.84% 和 95.77%。综上，基于 BGA-CRF 模型在 3 种模型中的良好表现，命名实体识别步骤选择采用此模型。

表 2 三种模型的评价指标

模型	精确率 P/%	召回率 R/%	F_1 值 /%
BiLSTM	93.13	93.69	93.16
BERT-BiLSTM-CRF	94.25	94.39	94.59
BGA-CRF	96.93	96.84	95.77

基于 Neo4j 道路交通事故领域风险因素知识图谱可视化 Neo4j 图数据库以其可扩展性和适应性著称[10]，具有实时更新、查询和编辑时不干扰现有数据的常规操作。相同实体的节点采用一致的风格表示，例如，以紫色表示事故 ID 实体的节点，而事故位置实体的节点则以橙色表示。棕色节点则表示这两种实体之间的关系，如“Accident Location”。该知识图谱的局部示意图如图 7 所示。

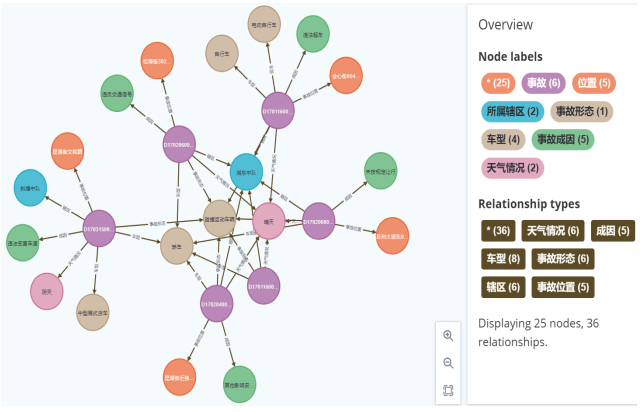

图 7 部分道路交通事故领域知识图谱示意图

3 知识图谱事故画像说明

在构建道路交通事故领域风险因素知识图谱的过程中，利用该图谱中的事故画像来详细描包括事故发生的时间、位置、天气情况、事故形态以及事故成因等信息。通过这种方式，知识图谱能够为每个交通事故提供一个全面而详细的视图，有助于深入理解和分析事故的各个方面。以某事故说明为例如图 8 所示。

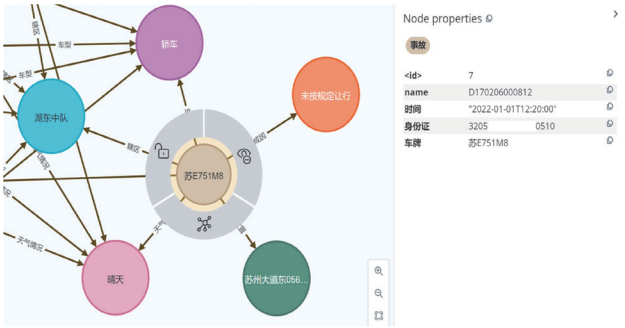

图 8 基于道路交通事故知识图谱的随机事故说明查询

苏 E751M8 车主在 2022 年 1 月 1 日 12 时 20 分驾驶轿车在湖东中队辖区苏州大道东 222 号路灯杆附近路段处因未按规定让行的行为，承担本次事故的主要责任。从当事人的事故数据中，可以看到事故代码、事故时间、车型、事故发生的位置、天气情况、所属辖区等信息。

4 问答系统

知识图谱问答系统的发展，结合人工智能、自然语言处理、图数据库和机器学习等多个领域的最新技术，为用户带来了更加智能和便捷的信息检索体验。以交通事故风险因素知识图谱为基础的智能问答系统如图 9 所示。

图 9 交通事故风险因素知识图谱问答系统界面

5 结语

本文通过对交通事故数据的系统化处理，提取了事故的关键信息，实现了数据从非结构化到结构化的转变。为了更直观地分析这些数据，采用 Neo4j 图数据库技术对数据进行存储与查询，构建道路交通事故数据风险因素知识库，不仅提升了数据的可视化分析能力，还增强了驾驶员对于驾驶时可能产生的道路风险因素的认知。此外，通过构建基于交通事故风险因素知识图谱测智能问答系统，提升了检索效率，实现了数据资源的有效共享，从而扩展了知识图谱的应用范围。

参考文献：

[1] 公安部办公厅统计处 .2023 年全国机动车和驾驶人统计分析 [J]. 公安研究 , 2024(4): 127-128.

[2] 高文馨 , 李贯峰 , 王云丽 , 等 . 融入逻辑规则的知识图谱推荐模型研究 [J]. 计算机技术与发展 , 2024, 34(9):109-115.

[3] 于德新 , 彭万里 , 吴新程 , 等 . 知识图谱和表示学习在道路交通事故数据挖掘中的应用 [J]. 安全与环境学报 , 2024, 24(10): 3950-3958.

[4] 王俞涵 , 陈子阳 , 赵翔 , 等 . 时序知识图谱表示与推理的研究进展与趋势 [J]. 软件学报 , 2024, 35(8): 3923-3951.

[5] 余礼根 , 郭晓利 , 赵红涛 , 等 . 基于 BERT-BiLSTM-CRF 模型的畜禽疫病文本分词研究 [J]. 农业机械学报 , 2024, 55(2): 287-294.

[6] 丁建平 , 李卫军 , 刘雪洋 , 等 . 命名实体识别研究综述 [J]. 计算机工程与科学 , 2024, 46(7): 1296-1310.

[7] 李浩君 , 方璇 , 戴海容 . 基于自注意力机制和双向 GRU 神经网络的深度知识追踪优化模型 [J]. 计算机应用研究 , 2022, 39(3): 732-738.

[8] 徐浩 , 廖铭新 , 吕家树 , 等 . 基于机器学习的工作井开挖周边管线沉降预测研究 [J]. 广东土木与建筑 , 2024, 31(6): 1-5.

[9] 杨文忠 , 丁甜甜 , 康鹏 , 等 . 基于舆情新闻的中文关键词抽取综述 [J]. 计算机工程 , 2023, 49(3): 1-17.

[10] 史政一 , 吕君可 , 黄弘 . 基于 Neo4j 的城市地下管道信息知识图谱构建研究 [J]. 中国安全生产科学技术 , 2024, 20(6): 5-10.

【作者简介】

张丽岩 (1978—)，女，黑龙江齐齐哈尔人，博士，高级实验师，研究方向：交通大数据、交通规划。

顾欣怡 (2000—)，女，江苏苏州人，硕士研究生，研究方向：交通大数据、交通规划。

马健 (1979—)，通信作者（email: 9764634@ qq.com），男，江苏扬州人，博士，副教授，研究方向：交通大数据、交通仿真与控制。

（收稿日期：2024-10-25）