

**学士学位论文中期报告**

**题目：基于知识图谱的交通肇事案件量刑推荐技术**

|  |  |
| --- | --- |
| **专　　业**： | 计算机科学与技术 |
| **学生姓名**： | **柳家成** |
| **学　　号**： | **18373111** |
| **指导教师**： | **阮利** |

**北京航空航天大学计算机学院**

2022年4月13日

**目　录**

[一、 课题背景 1](#_Toc13818)

[二、 研究现状 2](#_Toc13085)

[三、 研究内容与进展 2](#_Toc18466)

[3.1 知识图谱的构建 2](#_Toc19363)

[3.2 交通肇事案件判决书实体识别 3](#_Toc5215)

[四、 存在的问题 5](#_Toc11058)

[4.1 缺少交通肇事案件数据集 5](#_Toc30565)

[4.2 部分噪声数据的消除 5](#_Toc17709)

[五、 下阶段工作安排 6](#_Toc20960)

[六、 参考文献 6](#_Toc25108)

# 课题背景

随着“智慧司法”建设的不断推进，地区检查机关和相关业务部门已经实现线上办公，大幅提升了业务办理效率，与此同时如何通过信息技术对司法案件进行整理和分析成为研究的热点。

在2019年的余金平案件中检查院给出的缓刑建议和法院一审给出的有期徒刑两年和二审给出的有期徒刑三年六个月的结果相去甚远，引发巨大争议。在交通肇事案件的量刑方面，对量刑产生影响的有3个因素：被告人所触犯罪行，例如交通肇事罪、危险驾驶罪等；被告人犯罪的情节，例如肇事后逃逸、自首等；法官个人因素。在目前理想化的量刑流程中希望将法官个人的影响降低。而在2015年的一项研究中作者调研了数百份交通案件判决书，得出检察院给出的量刑建议被法院所接受的概率仅仅为57.9%的结论，说明在交通肇事量刑方面，由于各种证据错综复杂，法官个人因素对于量刑结果影响较大，本文试图使用知识图谱分析过往交通肇事相关的判决书，提取其中的相关信息，并能够根据输入的相关要素给出类似案件的推送以及量刑预测。

# 研究现状

目前知识图谱在司法领域得到广泛应用，但是在交通肇事领域的应用较少，在2015年丁琪等人手动分析了198份交通肇事判决书，分析了各种情况下检察院量刑建议与法院量刑结果的一致性，提出了量刑证据在司法适用上技术和制度的不足。

在量刑预测方面王治政、王雷等人提出使用向量表示实体和关系，并通过案件要素进行预训练得到量刑预测的任务，但是该方法的适用范围较小，仅仅分析了贩毒相关案件且准确率较低，在大部分测试集中准确率仅为20%左右，系统作出量刑预测的意义不是很大。黄治纲,谢新强等人提出以知识图谱为辅助信息,利用文本分类和信息抽取技术构建面向刑事案例的知识图谱,针对当事人的陈词供述,利用知识表示学习求解相似的案件，进一步实现法条推荐。但是在交通肇事领域应该更加关注法院根据被告人的犯罪事实来进行量刑，针对陈词供述所能提取到的信息有限。

# 研究内容与进展

本课题主要工作内容为对于知识图谱进行信息抽取和量刑预测与类案推荐两方面，目前完成了信息抽取中的知识图谱构建以及实体抽取的部分工作。

## 知识图谱的构建

在调研数百篇交通肇事案件判决书，以及对交警进行相关咨询后，整理出了与判决相关的部分结构：1：被告人所犯罪行，有无恶劣情节以及其所负有责任。2：与量刑相关情节，例如自首、赔偿等。3：根据中华人民共和国XX法XX条做出判决。其中被告人所触犯罪行一般分为危险驾驶罪和交通肇事罪，对量刑结果的影响最大，交通肇事罪量刑一般为有期徒刑，危险驾驶罪一般量刑一般为拘役和罚款。其次是恶劣情节（一般为逃逸），再次为自首、赔偿等相关情节。据此本文将知识图谱中的实体以及实体间关系定义如下。

**表1：知识图谱实体定义**

|  |  |
| --- | --- |
| 实体名 | 举例说明 |
| 被告人（person） | 被告人林XX |
| 罪名（charge） | 交通肇事罪 |
| 量刑情节（circumstance） | 自首、主动承认罪行等 |
| 拘役/有期徒刑（sentence1） | 拘役/有期徒刑XX年 |
| 缓刑（sentence2） | 缓刑XX年 |
| 罚金（sentence3） | 罚款XXX元 |
| 案件责任（responsibility） | 全部责任、部分责任等 |

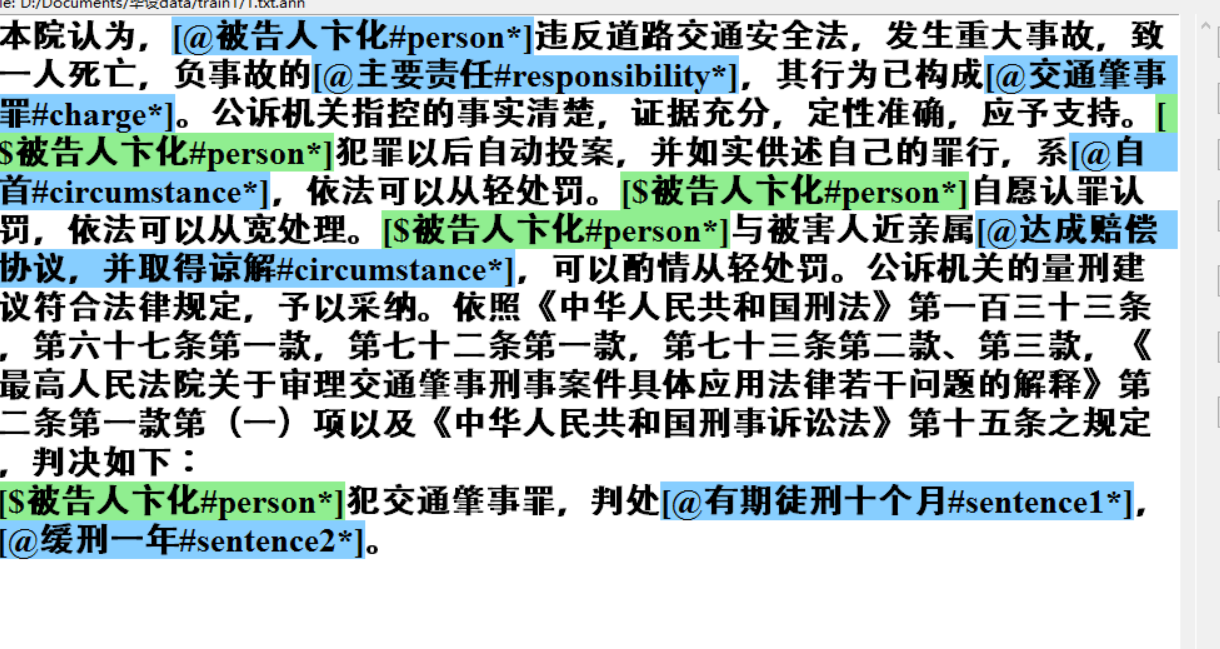
**表2：知识图谱关系定义**

|  |  |  |
| --- | --- | --- |
| 头实体 | 尾实体 | 关系 |
| 被告人 | 罪名 | 触犯 |
| 被告人 | 量刑情节 | 产生 |
| 拘役/有期徒刑 | 被告人 | 判处 |
| 缓刑 | 被告人 | 判处 |
| 罚金 | 被告人 | 判处 |
| 被告人 | 案件责任 | 负有 |

## 交通肇事案件判决书实体识别

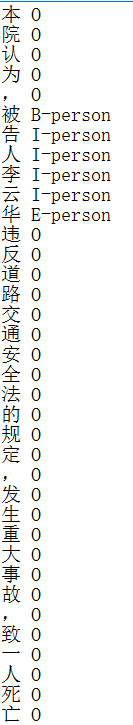
## 3.2.1数据集的构造

### 对于交通肇事相关裁判文书的判决部分进行标注处理，将其中的实体标注出来，在本文中使用手动标注工具YEDDA进行了标注，目前标注的裁判文书数量为100篇，其中交通肇事罪50篇，危险驾驶罪50篇。本文计划标注200篇裁判文书，其中160篇用于训练，40篇用于测试。



**图1：实体标注示例**

在手动完成标注后通过编程实现正则表达式匹配匹配出标注为实体的字段，并将YEDDA标注后的数据转换为BIOE格式。



**图2：转换为BIOE格式**

## 3.2.2基于LSTM模型与条件随机场的标签预测

在交通判决书的实体识别任务中上下文信息对于实体的识别相当重要，LSTM作为一种特殊的循环神经网络可以用来学习上下文字符的隐藏表示，在将数据集中的中文字符转换为字向量后进入双向LSTM模型并学习到字符和标注之间的函数关系，在处理子向量时就会给出该子向量的各种标签的预测分数。

在本文提出的交通肇事案件判决书中存在着诸如实体开头必为B-label、一个实体的所有组成部分的标签应该相同等隐性规则。条件随机场可以通过特征函数表达观测之间的依赖与前后多个状态之间的复杂依赖。首先需要定义一个特征函数集将标注序列作为输入，提取特征作为输出然后使用对数线性模型，其中s'为所有可能的状态序列，w为每个特征函数的权重，其对数似然函数正则化形式为。最优参数w为。训练结束后最优的状态序列为。在解码时使用维特比算法，从开始状态后每走一步，记录到达该状态所有路径的最大概率值，最后以最大值为基准继续向后推进，最后再从结尾回溯最大概率，完成实体的标注任务。

## 3.2.2基于远程监督的关系抽取

在交通肇事案件中实体之间的关系较为明确，例如“被告人”实体与“罪行”实体之间一定存在“犯下”关系。因此基于远程监督的关系抽思想非常适用于本场景。对于在完成实体识别后的交通肇事数据进行预处理，将每个字符转为字向量与两个实体，位置向量的结合体。然后进入卷积层使用三个卷积核进行卷积，得到3个向量。将每个向量按照两个实体位置划分为三段，分别进入池化层。将池化层的输出结果输入softmax层计算属于每种关系的概率值。

# 存在的问题

1. **缺少缺少人脸人脸**

## 缺少交通肇事案件数据集

目前数据集的来源为中国裁判文书网上的判决文书，用于模型训练时需要对其进行人工标注，工作较为繁琐，且数据量不足。

## 部分噪声数据的消除

在部分判决文书中出现了文书中所提到的证据并不能支撑量刑结果的情况，存在部分噪声无法消除，在总体数据量较小的情况下噪声的影响可能更大。

# 下阶段工作安排

|  |  |
| --- | --- |
| **时间期限** | **完成内容** |
| 2018.4.14 — 2016.4.18 | 完成剩下实体识别模型的数据集标注和训练 |
| 2016.4.19 — 2016.4.30 | 完成交通肇事案件三元组的抽取 |
| 2016.5.1 — 2016.5.10 | 完成类案推荐与量刑预测系统 |
| 2016.5.20 — 2016.5.31 | 撰写、修改毕业论文 |

# **参考文献**

1. 王治政,王雷,李帅驰,孙媛媛,陈彦光,许策,王刚,林鸿飞.基于多视角知识图谱嵌入的量刑预测[J].模式识别与人工智能,2021,34(07):655-665.DOI:10.16451/j.cnki.issn1003-6059.202107007.
2. 黄治纲,谢新强,邢铁军,葛东,蔡晨秋,窦丽莉,王天翊.基于司法案例知识图谱的类案推荐[J].南京大学学报(自然科学),2021,57(06):1053-1063.DOI:10.13232/j.cnki.jnju.2021.06.01
3. 袁珍珍. 基于民事案件法律文书的命名实体识别研究[D].中央民族大学,2021.DOI:10.27667/d.cnki.gzymu.2021.000506.
4. 刘明伟. 司法知识抽取与融合技术研究[D].中国电子科技集团公司电子科学研究院,2021.DOI:10.27728/d.cnki.gdzkx.2021.000137.
5. 陈彦光,王雷,孙媛媛,王治政,张书晨.面向法律文本的三元组抽取模型[J].计算机工程,2021,47(05):277-284.DOI:10.19678/j.issn.1000-3428.0057677.
6. 丁琪,陈颖,廖思谦.量刑证据的司法适用——以交通肇事罪判决书为样本的分析[J].河南警察学院学报,2015,24(02):96-100.DOI:10.16231/j.cnki.jhpc.2015.02.012.
7. Mintz M , Bills S , Snow R , et al. Distant supervision for relation extraction without labeled data[C]// ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore. Association for Computational Linguistics, 2009.
8. D Zeng, Kang L , Chen Y , et al. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks[C]// Conference on Empirical Methods in Natural Language Processing. 2015.
9. Lin Y , Shen S , Liu Z , et al. Neural Relation Extraction with Selective Attention over Instances[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016.
10. 李琳,解书贵,能沛然,陶晓辉.面向法律文书基于语句选择的证据预测方法[J].计算机与数字工程,2021,49(12):2437-2443.