基于司法案例的知识图谱构建技术研究

陈彦光 刘海顺 李春楠 刘静 孙媛媛 杨亮 林鸿飞

大连理工大学计算机科学与技术学院
信息检索研究室

大连市人民检察院技术处

2018年8月16日

目录

1 背景介绍

2 研究方法

3 实验相关

4 总结&展望

第一部分

背景介绍

背景介绍

研究背景

- 近年来我国不断深入推进“智慧司法”建设
- 中国裁判文书网已收入文书总量达5050余万篇，成为全球最大的裁判文书公开平台

研究目的

- 构建面向司法领域的知识图谱
- 实现面向业务的裁判文书智能检索等实际应用

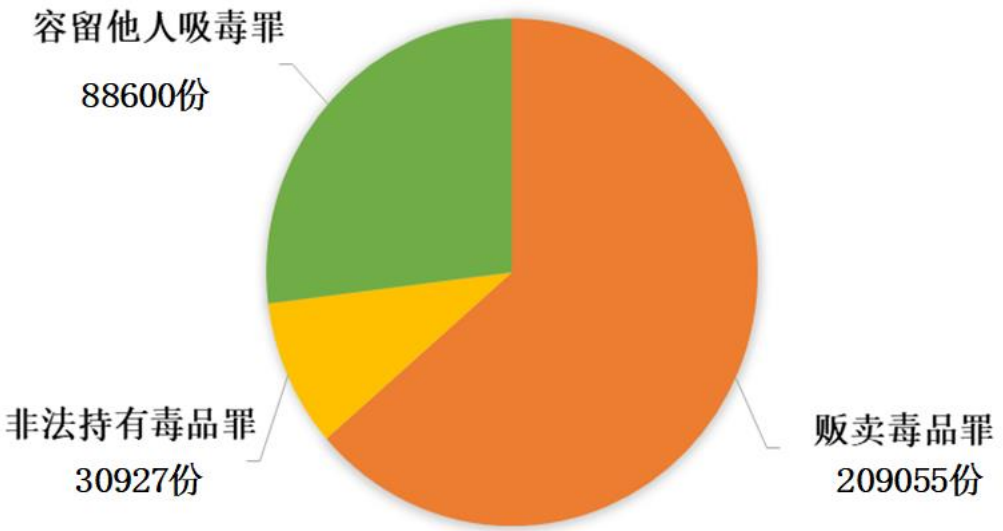
第二部分

研究方法

数据介绍

裁判文书网上公开的涉毒类案件刑事判决书共328582份

- 三类罪名：贩卖毒品罪、非法持有毒品罪、容留他人吸毒罪
- 简单案件：单人单情节案件 复杂案件：多人案件、多情节案件



研究数据

简单案例

李某某非法持有毒品案

江苏省常州经济开发区人民法院
刑事判决书
(2017)苏0492刑初235号

公诉机关江苏省常州经济开发区人民检察院。

被告人李某某。曾因吸毒、非法持有管制器具，分别于2013年8月28日被常州市武进区公安局处行政拘留十五日、十日；因吸毒，于2014年9月22日被常州市武进区公安局处行政拘留十五日后转社区戒毒三年、于2017年3月17日被常州市公安局常州经济开发区分局处行政拘留十五日后转强制隔离戒毒二年。因涉嫌犯非法持有毒品罪，于2017年5月8日被刑事拘留，5月18日被逮捕。

常州经济开发区人民检察院以常经检诉刑诉〔2017〕244号起诉书指控被告人李某某犯非法持有毒品罪，于2017年6月21日向本院提起公诉。本院立案受理后，依法适用简易程序，实行独任审判，公开开庭审理了本案。常州经济开发区人民检察院指派检察员潘剑云出庭支持公诉，被告人李某某到庭参加了诉讼。现已审理终结。

公诉机关指控，2017年3月17日，公安机关根据线索，在常州市武进区遥观镇宋剑湖家园107幢丙单元101室抓获涉嫌吸毒的被告人李某某，在其房间内查获疑似毒品7包、冰壶2个；2017年4月19日，公安机关在常州市武进区遥观镇宋剑湖家园107幢丙单元102室车库内又查获李某某藏匿于此的疑似毒品1包。经检验，上述疑似毒品中均检出甲基苯丙胺（冰毒）成份，净重合计10.76克。案发后，上述甲基苯丙胺10.76克、冰壶2个已由公安机关收缴。

归案后，被告人李某某如实供述了上述犯罪事实。

上述事实，被告人李某某在开庭审理过程中亦无异议，且有检察机关提交并经法庭质证、认证的下列证据予以证实：物证毒品、冰壶照片；书证常住人口信息表、称重记录及照片、保全决定书、保全物品清单、收缴毒品专用收据、行政处罚决定书、社区戒毒决定书、强制隔离戒毒决定书；证人杨某、李某、姚某的证言笔录；丁堰派出所民警出具的发破案及抓获经过、情况说明；检查笔录；尿液采集笔录、现场检测报告书；检验报告书。

复杂案例

车牌照，也没有汽车钥匙，就问靳车是怎么来的，靳某某告诉其是偷来的。因为其想要这辆车，就和靳商量以18000元的价格买下，如果不想要了再退钱。当天，其感觉是偷来的车，没有办法开，就联系和发短信让靳某某退车，靳不同意退车，其就把车开到了外甥女家里。

以上证据经当庭质证，上述证据能形成完整的证据链，予以确认。

二、非法持有毒品的犯罪事实

1、2014年2月初，被告人孟某某在长治市十中附近淘园村一游戏厅内，从一个不认识的人手中以每克100元的价格购买60克甲基苯丙胺（冰毒）供自己吸食。2014年3月2日，公安人员在其住处卧室内棕色布娄内当场查获48.9克可疑毒品。

经鉴定，从查获的48.9克可疑毒品中检出甲基苯丙胺。

以上犯罪事实公诉机关当庭出示的证据有：

（1）搜查笔录，证实2014年3月2日，晋城市公安局城区分局刑侦人员在长治市公安局刑侦人员的配合下，在长治市城区淮海中学附近一民房孟某某家中卧室内棕色布娄内发现一包可疑白色晶体块状物品，经称重为50.6克。

（2）晋城市公安局城区分局扣押物品清单及照片说明，证实查获在孟某某住处卧室内棕色布娄内查获毒品50.6克予以扣押。

（3）晋城市公安局城区分局现场检测报告书及照片说明，证实孟某某的检测样本经现场检测，结果呈阳性。

（4）晋城市公安司法鉴定中心出具的理化检验报告，证实从孟某某家中查获的可疑白色晶体包装，去包装后净重为48.94克，检出甲基苯丙胺成分。

（5）被告人孟某某在公安机关的供述，证实公安人员在其家中发现透明塑料袋装有的白色晶体块状物是冰毒，是其过年时在长治市十中附近淘园村一个游戏厅内和一个不认识的人购买的毒品冰毒，当时是每克100元的价格购买了60克，供其吸食。

以上证据经当庭质证，上述证据能形成完整的证据锁链，予以确认。

被告人孟庆伟辩护人当庭出示两份证据：1、孟××解除强制戒毒证明书（复印件），欲证明孟××（孟庆伟之父）系吸毒人员；2、死亡证明（复印件），欲证明孟××于2015年7月28日死亡。

公诉机关当庭质证意见认为：该证据从形式上系复印件，与本案缺乏关联性。

原审法院认为，该证据从证据形式上看均系复印件，未提供原件，不符合证据要件且不能证明查获的毒品系孟××所购，该证据不予确认。

2、2014年2月底，被告人陈×在长治市捉马村飞龙小区附近，从一个叫“阿伟”的湖南人手中，以每克100元的价格购买30克甲基苯丙胺（冰毒）供自己吸食。2014年3月3日，公安人员在其住处客厅茶几一铁盒内当场查获26.6克可疑毒品。

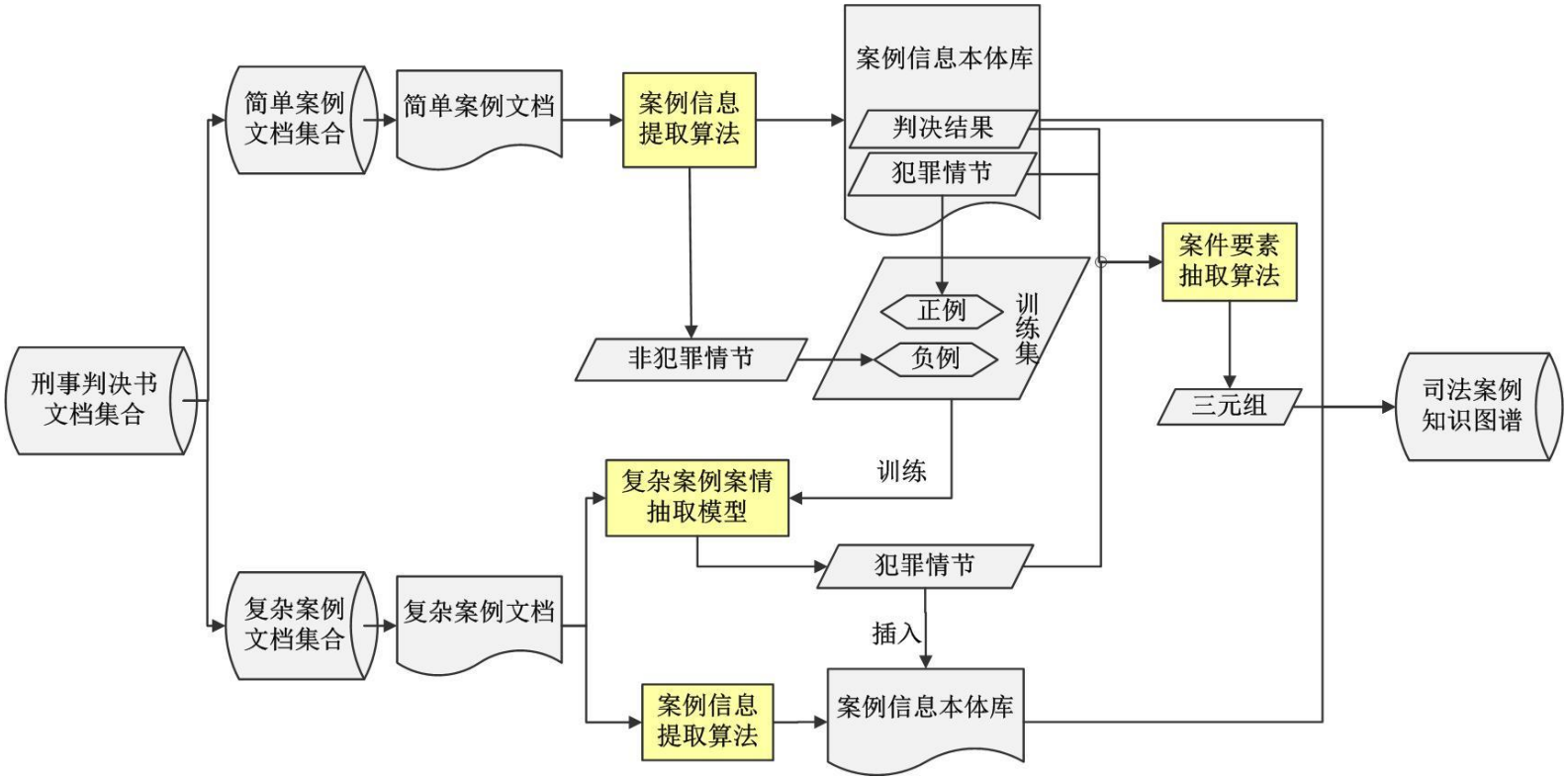
经鉴定，从查获的26.6克可疑毒品中检出甲基苯丙胺。

以上犯罪事实公诉机关当庭出示的证据有：

（1）搜查笔录，证实2014年3月3日，晋城市公安局城区分局刑侦人员在长治市公安局刑侦人员的配合下，在长治市城区西花园南区3号楼4单元802室陈×家中客厅茶几上一黄色铁盒内发现两包可疑白色晶体块状物品，经称重分别为10.4克、17.8克。

（2）晋城市公安局城区分局扣押物品清单及照片说明，证实查获在长治市城区西花

研究方法



- 案例信息提取算法
- 复杂案例案情抽取模型
- 案件要素抽取算法

案例信息提取算法

案例信息提取算法

案例本体结构

本体是对一个特定领域的重要概念的形式化描述。

案例本体结构	刑事判决书中对应内容
文书编号(FILEID)	(2014)云刑初字第1451号
刑事判决书标题(title)	陈某某贩卖毒品案
审判地点(place)	贵州省贵阳市云岩区人民法院
审判时间(time)	二〇一四年九月二日
公诉机关(prosecutor)	贵州省贵阳市云岩区人民检察院
被告人(defendant)	陈某某
犯罪类型(type)	贩卖毒品罪
犯罪情节(inf)	2014年4月24日12时许，被告人陈某某……所收缴的毒品为海洛因。
判决结果(judgement)	判处有期徒刑六个月，并处罚金人民币1000元。
审判依据(laws)	《中华人民共和国刑法》第三百四十七条……

案例信息提取算法

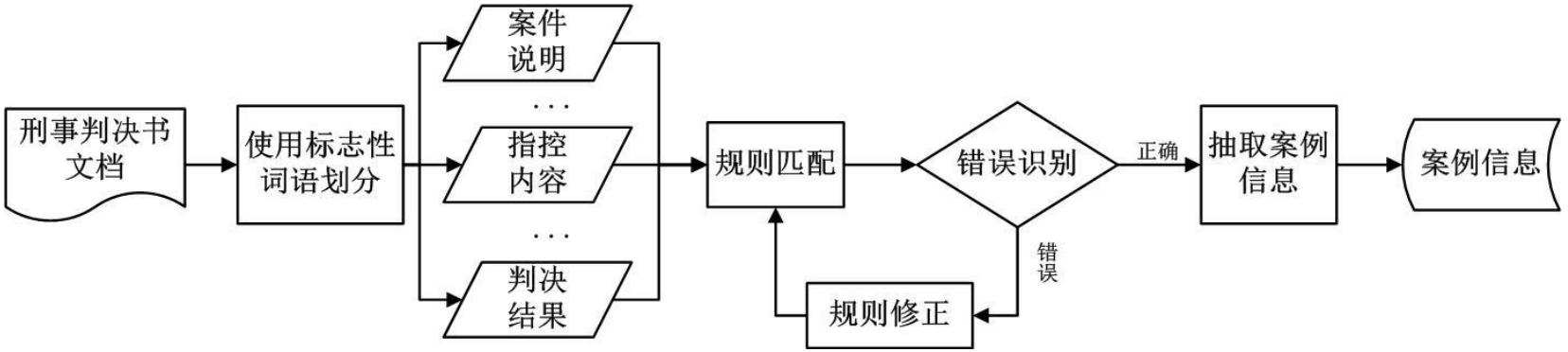
信息提取规则设计

1.刑事判决书的内容组织形式相对规范

2.在每个部分有标志性词语可作为信息提取的关键点

3.依照定义的案例本体结构对需要抽取的信息构造语法规则

案例信息提取流程



复杂案例案情抽取模型

复杂案例案情抽取模型

简单案例

李某某非法持有毒品案

江苏省常州经济开发区人民法院
刑事判决书
(2017)苏0492刑初235号

公诉机关江苏省常州经济开发区人民检察院。

被告人李某某。曾因吸毒、非法持有管制器具，分别于2013年8月28日被常州市武进区公安局处行政拘留十五日、十日；因吸毒，于2014年9月22日被常州市武进区公安局处行政拘留十五日后转社区戒毒三年、于2017年3月17日被常州市公安局常州经济开发区分局处行政拘留十五日后转强制隔离戒毒二年。因涉嫌犯非法持有毒品罪，于2017年5月8日被刑事拘留，5月18日被逮捕。

常州经济开发区人民检察院以常经检诉刑诉〔2017〕244号起诉书指控被告人李某某犯非法持有毒品罪，于2017年6月21日向本院提起公诉。本院立案受理后，依法适用简易程序，实行独任审判，公开开庭审理了本案。常州经济开发区人民检察院指派检察员潘剑云出庭支持公诉，被告人李某某到庭参加了诉讼。现已审理终结。

公诉机关指控，2017年3月17日，公安机关根据线索，在常州市武进区遥观镇宋剑湖家园107幢丙单元101室抓获涉嫌吸毒的被告人李某某，在其房间内查获疑似毒品7包、冰壶2个；2017年4月19日，公安机关在常州市武进区遥观镇宋剑湖家园107幢丙单元102室车库内又查获李某某藏匿于此的疑似毒品1包。经检验，上述疑似毒品中均检出甲基苯丙胺（冰毒）成份，净重合计10.76克。案发后，上述甲基苯丙胺10.76克、冰壶2个已由公安机关收缴。

归案后，被告人李某某如实供述了上述犯罪事实。

上述事实，被告人李某某在开庭审理过程中亦无异议，且有检察机关提交并经法庭质证、认证的下列证据予以证实：物证毒品、冰壶照片；书证常住人口信息表、称重记录及照片、保全决定书、保全物品清单、收缴毒品专用收据、行政处罚决定书、社区戒毒决定书、强制隔离戒毒决定书；证人杨某、李某、姚某的证言笔录；丁堰派出所民警出具的发破案及抓获经过、情况说明；检查笔录；尿液采集笔录、现场检测报告书；检验报告书。

复杂案例

车牌照，也没有汽车钥匙，就问靳车是怎么来的，靳某某告诉其是偷来的。因为其想要这辆车，就和靳商量以18000元的价格买下，如果不想要了再退钱。当天，其感觉是偷来的车，没有办法开，就联系和发短信让靳某某退车，靳不同意退车，其就把车开到了外甥女家里。

以上证据经当庭质证，上述证据能形成完整的证据链，予以确认。

二、非法持有毒品的犯罪事实

1、2014年2月初，被告人孟某某在长治市十中附近淘园村一游戏厅内，从一个不认识的人手中以每克100元的价格购买60克甲基苯丙胺（冰毒）供自己吸食。2014年3月2日，公安人员在其住处卧室内棕色布娄内当场查获48.9克可疑毒品。

经鉴定，从查获的48.9克可疑毒品中检出甲基苯丙胺。

以上犯罪事实公诉机关当庭出示的证据有：

（1）搜查笔录，证实2014年3月2日，晋城市公安局城区分局刑侦人员在长治市公安局刑侦人员的配合下，在长治市城区淮海中学附近一民房孟某某家中卧室内棕色布娄内发现一包可疑白色晶体块状物品，经称重为50.6克。

（2）晋城市公安局城区分局扣押物品清单及照片说明，证实查获在孟某某住处卧室内棕色布娄内查获毒品50.6克予以扣押。

（3）晋城市公安局城区分局现场检测报告书及照片说明，证实孟某某的检测样本经现场检测，结果呈阳性。

（4）晋城市公安司法鉴定中心出具的理化检验报告，证实从孟某某家中查获的可疑白色晶体包装，去包装后净重为48.94克，检出甲基苯丙胺成分。

（5）被告人孟某某在公安机关的供述，证实公安人员在其家中发现透明塑料袋装有的白色晶体块状物是冰毒，是其过年时在长治市十中附近淘园村一个游戏厅内和一个不认识的人购买的毒品冰毒，当时是每克100元的价格购买了60克，供其吸食。

以上证据经当庭质证，上述证据能形成完整的证据锁链，予以确认。

被告人孟庆伟辩护人当庭出示两份证据：1、孟××解除强制戒毒证明书（复印件），欲证明孟××（孟庆伟之父）系吸毒人员；2、死亡证明（复印件），欲证明孟××于2015年7月28日死亡。

公诉机关当庭质证意见认为：该证据从形式上系复印件，与本案缺乏关联性。

原审法院认为，该证据从证据形式上看均系复印件，未提供原件，不符合证据要件且不能证明查获的毒品系孟××所留，该证据不予确认。

2、2014年2月底，被告人陈×在长治市捉马村飞龙小区附近，从一个叫“阿伟”的湖南人手中，以每克100元的价格购买30克甲基苯丙胺（冰毒）供自己吸食。2014年3月3日，公安人员在其住处客厅茶几一铁盒内当场查获26.6克可疑毒品。

经鉴定，从查获的26.6克可疑毒品中检出甲基苯丙胺。

以上犯罪事实公诉机关当庭出示的证据有：

（1）搜查笔录，证实2014年3月3日，晋城市公安局城区分局刑侦人员在长治市公安局刑侦人员的配合下，在长治市城区西花园南区3号楼4单元802室陈×家中客厅茶几上一黄色铁盒内发现两包可疑白色晶体块状物品，经称重分别为10.4克、17.8克。

（2）晋城市公安局城区分局扣押物品清单及照片说明，证实查获在长治市城区西花

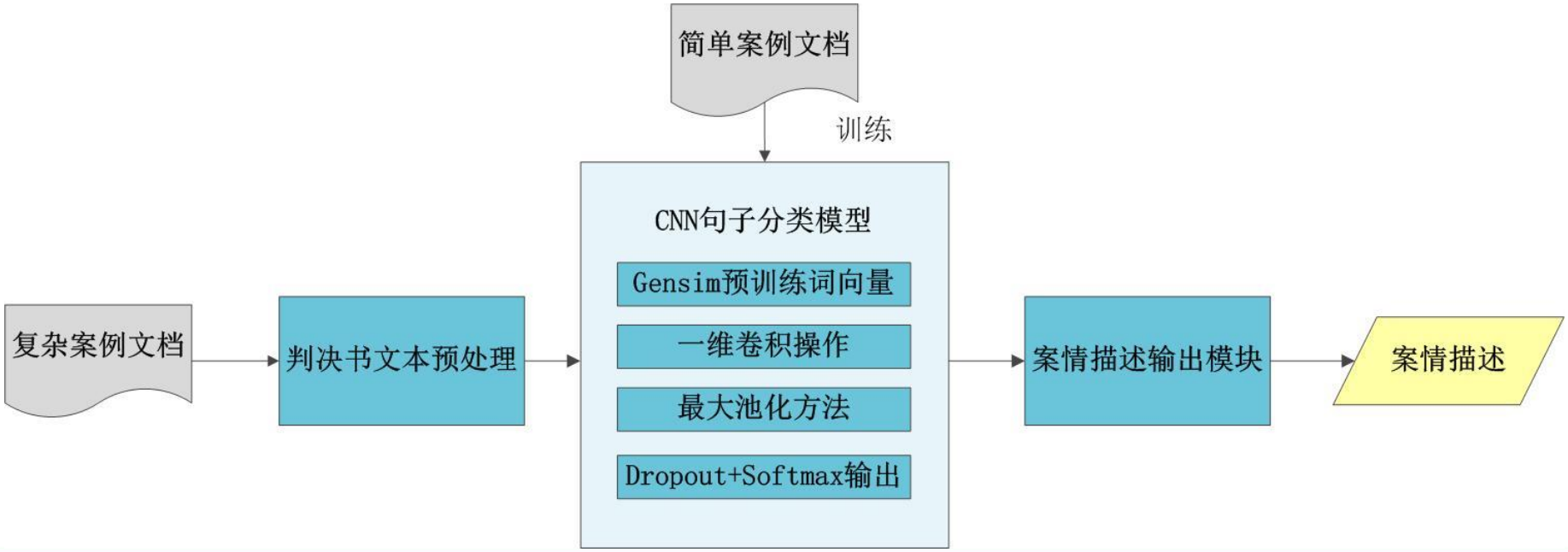
复杂案例案情抽取模型

基于卷积神经网络的案情描述句子分类模型

1.复杂案例文档中案情描述部分分散在不同段落，难以通过规则直接提取

2.缺少基于复杂案例的案情描述句和非案情描述句标注数据集

3.复杂案例的案情描述句和简单案例的案情描述句具有一致性



案件要素抽取算法

案件要素抽取算法

案件要素识别

1.使用哈工大的语言技术平台进行自然语言处理

2.结合涉毒类案件领域知识，对非法毒品名称、毒品重量等实体进行二次抽取

关系定义

犯罪关系	判决结果
卖	罚金
买	拘役
持有	有期徒刑
容留	无期徒刑
吸食	死刑

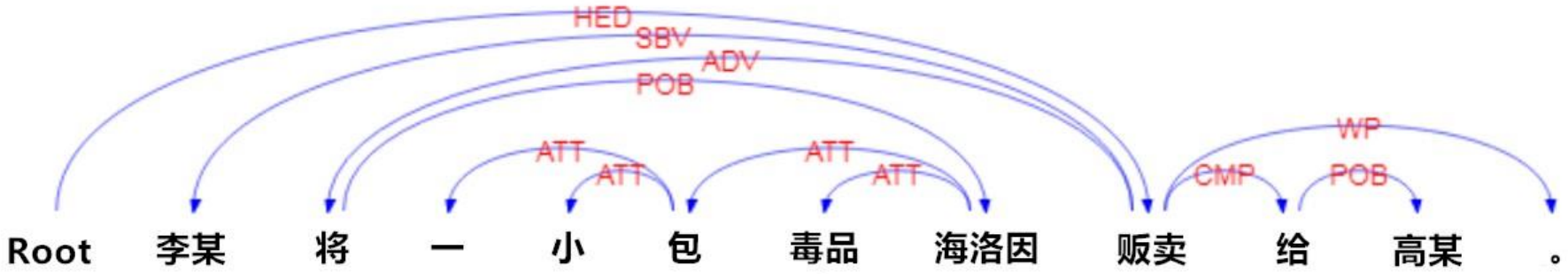
案件要素抽取算法

三元组构建和存储

1.对文本进行依存句法分析，确定语句中各要素之间的句法关系

2.将识别案件要素以三元组形式表示

3. 使用MySQL关系数据库存储知识图谱



第三部分

实验分析

数据集

裁判文书网上公开的涉毒类案件刑事判决书共328582份

- 三类罪名：贩卖毒品罪、非法持有毒品罪、容留他人吸毒罪

罪名	数量
贩卖毒品罪	209055
非法持有毒品罪	30927
容留他人吸毒罪	88600

实验结果与分析

案例信息提取实验

评价指标：

$$准确率 = \frac{正确提取信息的文档数}{进行信息提取的文档数}$$

$$召回率 = \frac{提取全部信息的文档数}{总文档数}$$

案件名称	准确率	召回率
贩卖毒品罪	80.15%	93.47%
非法持有毒品罪	82.34%	97.26%
容留他人吸毒罪	81.04%	94.94%

复杂案例的案情描述提取实验

训练集：随机选取简单案例的案情描述5000句为正例，非案情描述5000句作为负例

测试集：在复杂案例中按上述要求人工选取1000句构成测试集。

训练集大小	准确率
2000句	75.26%
10000句	91.51%

在10000句训练集的基础上实验结果：

文本分类方法	准确率
逻辑回归方法	84.15%
SVM	85.34%
随机森林	84.40%
Text CNN	**91.51%**

复杂案例的案情描述提取实验

针对三类案件分别训练模型进行情节提取实验

统计情节提取实验结果的准确率

案件类别	准确率
贩卖毒品罪	91.51%
非法持有毒品罪	93.24%
非法容留他人吸毒罪	89.77%

司法知识图谱构建

“实体-关系-实体”三元组共274万余个

包含涉及量刑的犯罪情节和判决结果的信息

案例本体	实例中对应内容
犯罪情节	1.2017年4月初的一天，被告人陈某在华蓥市XX路XX号其家中容留王某某吸食毒品甲基苯丙胺（冰毒）。 2.2017年4月20日下午，被告人陈某在华蓥市XX路XX号其家中容留王某吸食毒品甲基苯丙胺。 3.2017年4月21日下午，被告人陈某在华蓥市XX路XX号其家中容留张某、柏某吸食毒品甲基苯丙胺。
判决结果	判处有期徒刑九个月，并处罚金人民币6000元。

司法知识图谱构建

“实体-关系-实体”三元组共274万余个

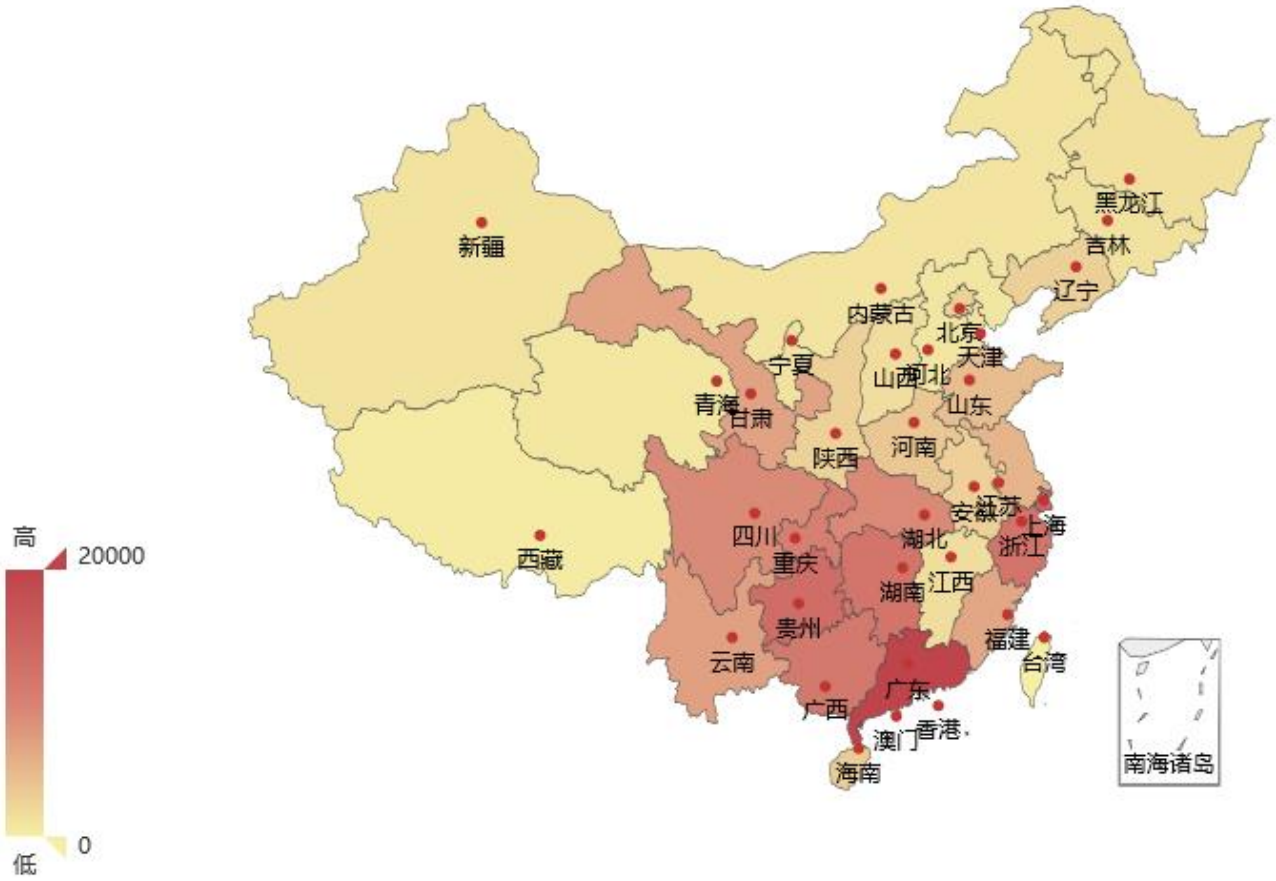
包含涉及量刑的犯罪情节和判决结果的信息

实体1	关系	实体2
陈某	容留	柏某
陈某	容留	张某
陈某	**有期徒刑**	**[0，9，0]**
柏某	吸食	冰毒
陈某	容留	王
陈某	吸食	甲基苯丙胺
陈某	**罚金**	6000
陈某	容留	王某
……		

刑期以[年,月,日]的形式存储

知识图谱应用

案件分布地域统计（贩卖毒品罪在各省份分布情况）



实验结果与分析

三类案件的罚金分布统计：

罚金等级	贩卖毒品罪	非法持有毒品罪	容留他人吸毒罪
<3000元	51.96%	27.16%	45.35%
3000元-6000元	30.09%	36.25%	38.48%
6000元-9000元	3.73%	5.25%	4.39%
>=9000元	14.22%	31.34%	11.78%

罚金与毒品克数的关联分析：

（贩卖甲基苯丙胺案件中罚金等级所占比例）

毒品克数	<3000元	3000元-6000元	>=6000元
<5克	47.30%	23.09%	29.61%
5克-20克	15.27%	25.08%	59.65%
20克	5.94%	15.92%	78.14%

第四部分

总 结

总 结

- 本文提出了一种半监督的方法，保证在算法复杂度不高的情况下处理海量的裁判文书，实现了在缺少标注数据的情况下，构建司法领域的知识图谱。
- 基于构建的知识图谱可实现对相关案件关键情节和判决结果的统计分析，为司法文书智能处理提供数据基础。

未来工作

- 对已构建的知识图谱进行扩充、完善和实体消岐
- 改进案件要素提取算法

致谢

本文的研究工作受最高人民检察院检察技术信息研究中心基本科研业务费专项资金课题项目和科技部“十三五”重点研发计划项目“智能辅助检察办案关键技术研究”资助。

主要参考文献

[1] Filtz E.: Building and Processing a Knowledge-Graph for Legal Data. In: European Semantic Web Conference, pp.184-194.Portoroz, Slovenia(2017).
[2] 何庆, 汤庸, 黄永钊.:基于本体的法律知识库的研究与实现.计算机科学 (02):175-177(2007).
[3] 佘贵清, 张永安.:审判案例自动抽取与标注模型研究.现代图书情报技术, (6):23-29(2013).
[4] Luo B, Feng Y, Xu J, et al.: Learning to Predict Charges for Criminal Cases with Legal Basis. In: Conference on Empirical Methods in Natural Language Processing, pp.2727-2736. Copenhagen, Denmark (2017).
[5] Kim Y. :Convolutional Neural Networks for Sentence Classification. Eprint Arxiv:(2014).
[6] 刘挺, 车万翔, 李正华.: 语言技术平台. 中文信息学报25(6):53-62(2011).

谢谢！