

法院判决书关键信息抽取系统设计与实现

刘 稳, 王 锦, 李 锐, 游景扬, 陈建峡

(湖北工业大学计算机学院, 湖北 武汉 430068)

[摘 要] 从海量的法院判决书数据中快速抽取关键信息, 构建结构化的数据, 对法院信息化建设具有重要的现实意义。为实现这一目标, 结合法院判决书文本的特点, 设计了法院案件信息抽取模型, 基于法院案件的命名实体识别、框架知识表示和事件信息抽取等关键技术, 研发了法院判决书信息抽取系统。实验结果表明, 该系统不仅能够自动生成结构化的数据信息, 而且结合主题图技术进行可视化展示, 以供用户快速查询和修改, 有助于案件判决相关人员提高犯罪信息分析的质量和效率。

[关键词] 命名实体识别; 框架知识表示; 信息抽取

[中图分类号] TP391

[文献标识码] A

信息抽取(Information Extraction)的主要功能是从文本中抽取特定的事实信息, 这些文本可以是结构化、半结构化或非结构化的数据。通常信息抽取利用机器学习、自然语言处理(Natural Language Processing, NLP)等方法从上述文本中抽取特定的信息后, 保存到结构化的数据库当中, 以使用户查询和使用^[1]。

目前, 法院案件判决领域逐步实现了信息化, 但仍存在大量的非结构化文本信息。面对日益增长的大量案件信息、涉案人员等信息数据, 案件裁决人员需要花费很多时间在阅读案件笔录上和 Related 历史案件的分析上。若能利用信息抽取技术, 将各类案件文本中的信息点分析出来, 对涉案人员、案情信息等进行智能化管理, 便于日后的查询和各层级法院之间的信息共享。

美国克莱蒙研究生院开发了一个自动的犯罪信息报导与调查访谈系统^[2-3]。该系统利用认知心理的访谈技术, 唤起证人的回忆信息, 用自然语言记录案件情况, 然后利用信息抽取技术, 从证人叙述与访谈对话记录中抽取犯罪相关实体。美国亚利桑那州大学利用知识库、机器学习和少量手工规则的方法, 对人名、住址、工具、麻醉药物、私人财物等实体进行识别和抽取^[4-5], 开发了基于神经网络的实体抽取系统。

国内在法律领域对基于数据库的构建和数据挖掘技术研究较多, 对自然语言文本进行信息抽取研究较少。文献[6]对公安案件文本领域词汇的获取、命名实体的识别、实体关系的抽取等模块进行了研究。文献[7]采用文本挖掘的相关技术, 主要实现了给定案件的相似性判别和文本聚类功能。然而, 采用信息抽取技术对法院文书的处理尚处于起步阶段。

本文综合应用已有的信息抽取技术, 自动对判决书中案件涉及到的人、时间、地点、事件等信息进行抽取, 不仅能够找出这些实体间的关系, 并通过可视化的方式展现出来, 以供查询和修改, 使案件判决相关人员免于阅读大量的案卷, 提高犯罪信息分析的质量和效率, 从而更快速准确地给出公平公正的裁决结果。

1 信息抽取关键技术简介

项目研究中的信息抽取技术主要包含命名实体识别研究、框架理论方法和事件抽取等关键技术研究。

命名实体识别(Named Entity Recognition, NER)是信息抽取工作中最基本的环节。它的任务是从文本中提取出具有特殊含义或信息的名词和短语, 并为之添加标注信息, 为后续的工作奠定基础。

[收稿日期] 2017-01-11

[基金项目] 湖北省教育厅青年基金(Q20141420)

[第一作者] 刘 稳(1995-), 男, 湖北天门人, 湖北工业大学本科生, 研究方向为机器学习

一般来说,命名实体识别主要是识别出文本的三大类(实体类、时间类和数字类)、七小类(人名、机构名、地名、时间、日期、货币和百分比)^[8]。

框架是 M.Minsky 在 1957 年提出的,最早用于视觉感知、自然语言对话等问题的表示,目前作为一种表示知识对象(实体)的数据结构。框架理论认为,人们对现实世界中的事物的认知都是以一种类似于框架的结构存储于记忆中的,当遇到新事物时,就倾向于从记忆中找到一种合适的框架,并根据实际情况对其做出适当的修改^[9]。

事件抽取是信息抽取一个重要的研究方向。它将事件提取为结构化的数据。ACE (Automatic Content Extraction) 将事件抽取的工作定义为检测与识别事件。即识别指定事件,并从中抽取特定信息。因此 ACE 将事件抽取分为事件类别识别和事件元素识别。

2 系统总体设计

2.1 系统总体设计

法院判决书信息抽取系统的总体流程见图 1。

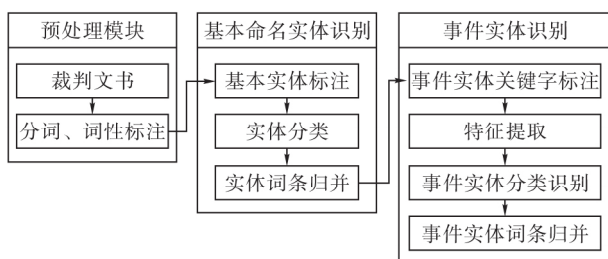


图 1 系统总体框架图

2.2 系统数据库设计

数据库中,与主题图相关的表有三张:case、case_people、case_casepeople。case 表中(图 2)存储案件信息和法院信息,如:案件细节、案件结果、法院名、法院等级;case_casepeople 表(图 3)存储涉案人员信息,如:姓名、性别、身份;case_people 表(图 4)不存储实质性信息,而是作为桥梁,将 case 表和 case_people 表连接起来。

id	case_name	court_name	court_date	case_detail	case_result	case_level
1	(2012)滨邹平县人	滨中	1 2012-11	3 行再关于	一、	才
2	(2012)滨山东省滨	滨中	2 2012-12	3 行再原审上	驳回上	
3	(2012)滨山东省滨	滨中	2 2013-10	3 行初曹福东	驳回原	
4	(2012)滨山东省滨	滨中	2 2013-10	3 行初曹福东	驳回原	

图 2 case 数据库表

id	case_id	case_people	case_type
1	1	1	1
2	1	2	1
3	1	3	2
4	1	4	2

图 3 case_casepeople 数据库表

id	name	type	gender	birthday	native_pl	work_planation
2	王刚	3	1	1977-12		汉族
5	李绪	1	2	1946-12		汉族
6	王刚	3	1	1977-12		汉族
9	曹福东	1	1	1969-12		汉族
15	曹福东	1	1	1969-12		汉族

图 4 case_casepeople 数据库表

抽取、CRF 学习得出 CRF 训练模型,在命名实体识别时,可以使用该 CRF 模型提高命名实体识别的准确率。具体流程见图 5。

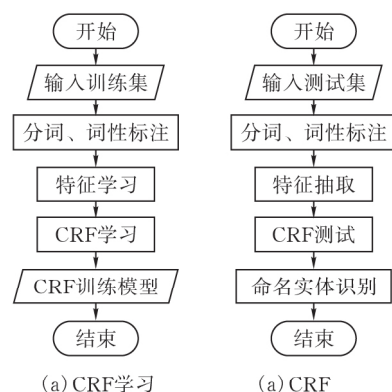


图 5 CRF 命名实体流程图

3.2 法院判决书框架知识表示

框架在文本信息进行事件抽取的应用十分广泛。其中,一个事件是由不同方面的子话题构成的,一个子话题是一个“事件侧面”,能发出“事件侧面”的词语称为“侧面词”。通过预先定义好的框架、事件侧面及侧面词,利用框架构造抽取规则和侧面词匹配规则抽取文本中的指定信息^[9]。

法院判决书的信息抽取需要从法院判决书中抽取出案件信息、法院信息和涉案人信息等三种关键信息。其中,案件信息应该有案件名、案件细节、审判结果等属性;法院信息应有法院名、法院等级等属性;涉案人信息应该包含姓名、性别、身份、民族等属性。用框架表示方法表示(图 6)。

将抽取案件信息分解为案件名、案件细节等内容的抽取。将案件名抽取转换为“侧面词”的匹配。用 CRF 模型判断匹配的结果是否正确。

3.3 法院判决书事件抽取

事件触发词引发事件的产生,是决定事件是否出现的重要特征。一般来说,每个事件都有特定的触发词。在专业领域相关性比较强、触发词的歧义性不大时,一般研究都根据触发词,用模式匹配的方法挑选出候选事件。再根据特定的算法判断候选事

框架名: <案件信息>	
案件名:	(2012)滨中民保字第 1 号-4
案件细节:	张三因不服法院裁决,向法院提出民事诉讼
案件结果:	驳回上诉
审判时间:	2011-5-8
框架名: <法院信息>	
法院名:	滨州中级人民法院
法院名缩写:	滨中
法院等级:	中级
框架名: <涉案人信息>	
姓名:	张三
性别:	男
出生日期:	1989 年 5 月 50 日
民族:	汉
籍贯:	湖北省武汉市
身份:	被告

图 6 法院判决书信息抽取框架表示

件的准确性。候选事件即含触发词的句子,这个句子可能是一个事件。

法院判决书事件抽取模块流程图见图 7。

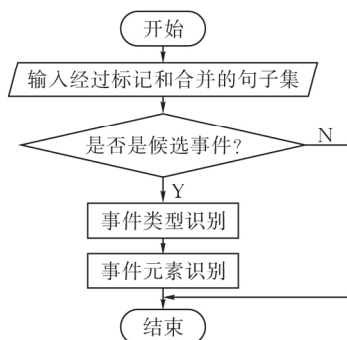


图 7 法院判决书事件抽取模块流程

3.3.1 基于触发词的候选事件获取 事件元素抽取即在确定了事件类型之后,从文本中抽取事件组成部分,并将其正确地分类。比如:

申请人:张玉华,男,1953 年 2 月 2 日出生,汉族,住沾化县大高镇大张村。

在确定了这个句子是一个涉案人信息事件后,如何从句子中抽取事件的组成部分;如何提取出“张玉华”;如何将“张玉华”判断为原告姓名。

根据参考文献[6-7,9],本文为案件信息事件、法院信息事件、涉案人员信息事件挑选出了触发词(表 1)。

表 1 触发词

事件类型	触发词
案件信息	案由、案件、本院依据
法院信息	法院、刑、行、赔、民、发、知
人员信息	申请人、被申请人、原告、被告

3.3.2 基于触发词的事件元素获取 根据知识框架理论,将案件信息、法院信息、涉案人信息看作框架,每个框架有各自的“槽”面,每个“槽”有独特的“侧面词”,具体情况见表 2。

表 2 法院判决书信息框架表示

框架	槽	侧面词
案件信息	案件细节	现因、本预案根据
	案件结果	裁定如下、合议庭意见
	审判日期	……年……月……日
	案件名	对……案
法院信息	法院名	法院
	法院登记名称缩写	初级、中级、高级、最高第
人员信息	姓名	原告:、被告:、申请人:、被申请人:
	性别	性
	出生日期	出生
	民族	族
	籍贯	出生于……省……市
	身份	原告、被告

4 系统实现

系统界面和功能模块实现见图 8、图 9、图 10。



图 8 系统界面图

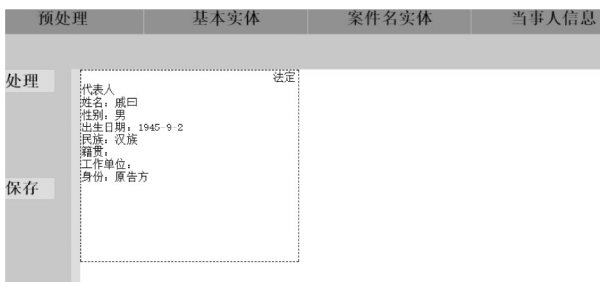


图 9 系统预处理图

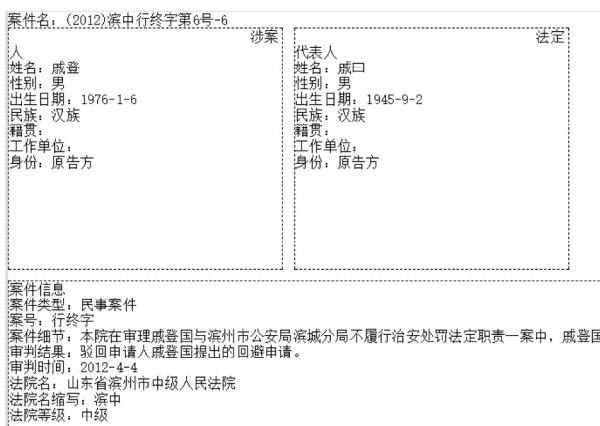


图 10 系统处理结果图

5 系统测试

5.1 实验结果

客户提供了 17858 篇判决书,其中 16281 篇是可以读取的,实验结果见表 3、表 4。

表 3 法院判决书事件抽取结果表示

事件类型	案件文本数	候选事件数	事件数
案件信息	16 281	19 283	16 281
法院信息	16 281	15 792	14 489
人员信息	16 281	89 631	84 097

表 4 法院判决书事件元素抽取结果表示

框架	槽	候选数	准确数
案件信息	案件细节	17 530	16 983
	案件结果	17 149	15 469
	审判日期	18 554	18 142
	案件名	18 962	18 763
法院信息	法院名	15132	13115
	法院等级	13 907	13 124
	名称缩写	13 086	11 953
人员信息	姓名	87 636	76 409
	性别	75 910	64 862
	出生日期	36 985	34 326
	民族	48 043	45 641
	籍贯	50 632	49 806
	身份	85 073	80 749

5.2 结果评测

用准确率(P)、召回率(R)和调和平均值(F)来评测实验结果。计算公式如下^[9]:

$$P = \frac{\text{系统显示正确的信息点}}{\text{系统显示的所有信息点}}$$

$$R = \frac{\text{系统显示正确的信息点}}{\text{测试集中出现的所有信息点}}$$

$$F = 2 \times P \times R / (P + R)$$

根据实验结果,系统的测评结果见表 5。

表 5 法院判决书事件抽取结果评测表示

	准确率 $P/\%$	召回率 $R/\%$	调和平均值 $F/\%$
案件信息	94.56	84.52	89.66
法院信息	90.16	91.75	90.95
人员信息	89.64	93.83	91.69

从整体上看,抽取结果准确率还是令人满意的,

但是案件信息的召回率有点低。

6 总结与展望

本文设计了法院案件信息抽取模型,基于法院案件的命名实体识别、框架知识表示和事件信息抽取等关键技术,实现了法院判决书信息抽取系统。实验结果表明,该系统有助于案件判决相关人员提高犯罪信息分析的质量和效率。

本文仅对基于语义相似度的计算作了初浅的研究,未来对于语义理解的相似度计算必然成为中文文本处理的主流,因为这种方法更适合汉语语言的特点和习惯。如何建立一个更好的语义理解模型,把它应用到更多的具体领域进行验证,如聚类、自动文摘等,是下一步研究的重点。

[参 考 文 献]

- [1] 李保利,陈玉忠,俞士汶. 信息抽取研究综述[J]. 计算机工程与应用, 2003, 39(10):1-5.
- [2] Ku C H, Iriberry A, Leroy G. Crime information extraction from police and witness narrative reports[J]. Human Biology, 2008, 80(6):593-600.
- [3] Ku C H, Iriberry A, Leroy G. Natural language processing and e-Government: crime information extraction from heterogeneous data sources[C]// International Conference on Digital Government Research. Digital Government Society of North America, 2008: 162-170.
- [4] Chen H, Chung W, Xu J J, et al. Crime data mining: a general framework and some examples[J]. Computer, 2004, 37(4):50-56.
- [5] Chau M, Xu J J, Chen H. Extracting meaningful entities from police narrative reports[C]// Proceedings of the National Conference for Digital Government Research, 2002: 271-275. 129 Los Angeles, California; Digital Government Research Center, 2002.
- [6] 乔春庚. 公安领域案件信息抽取系统设计与实现[D]. 北京:北京机械工业学院,北京信息科技大学, 2007.
- [7] 徐亚娟. 基于公安业务信息的文本挖掘技术研究与应用[D]. 杭州:浙江大学, 2008.
- [8] 郭喜跃,何婷婷. 信息抽取研究综述[J]. 计算机科学, 2015, 42(2):14-17.
- [9] 陈慧炜. 刑事案件文本信息抽取研究[D]. 南京:南京师范大学, 2011.

Design and Implementation of Key Information Extraction System in Court Verdict

LIU Wen, WANG Jin, LI Rui, YOU Jingyang, CHEN Jianxia

(School of Computer Science, Hubei Univ. of Tech., Wuhan, 430068, China)

Abstract: The rapid extraction of key information and the construction of structured data from the mass court judgment data have important practical significance for the court's information construction. According to the characteristics of court verdict, the paper designs the information extraction model of court cases, and realizes the information extraction system based on the named entity recognition, framework knowledge representation and event information extraction. The experimental results show that the proposed system can not only generate the structured data automatically, but display visualizations of the data conveniently by using the thematic map technology to help users query and modify quickly, which can help the judgment of the involved personnel to improve the quality and efficiency of the criminal information analysis.

Keywords: named entity recognition; framework knowledge representation; information extraction

[责任编辑: 张岩芳]

(上接第 62 页)

Multi-Exposure Image Fusion Based on Luminance Consistency

WANG Shuqing, LI Yewei

(School of Electrical and Electronic Engin., Hubei Univ. of Tech., Wuhan 430068, China)

Abstract: A new multi-exposure image fusion algorithm is proposed to address the problem of the local luminance is inconsistent with the corresponding scene in traditional multi-exposure images. By analyzing the multi-exposure image sequence, the algorithm decomposes each image into three conceptually independent components: local contrast, structure information and exposure luminance information. On the one hand, it retains more details of the scene information and scene structure information; on the other hand, it keeps the fused image local luminance consistent with the corresponding scene luminance. Eight sets of multi-exposure image sequence were tested in the experiments, and the result was compared with other seven kinds of traditional methods. The experimental results demonstrate that the fusion image is rich in detail with good structural similarity and luminance consistency, showing a natural and vivid visual effect.

Keywords: multi-exposure image fusion; luminance consistency; structure information; local contrast

[责任编辑: 张岩芳]