公安领域案件文本信息抽取研究综述

陈慧炜

(南京师范大学 文学院,江苏 南京 210097)

摘 要:公安领域存在大量非结构化案件文本,使人工查询与整理存有困难。信息抽取作为应对海量信息的一门技术,能够有效处理案件信息的结构化问题。本文总结了目前信息抽取的方法技术,在实体识别、触发词获取和事件抽取等子任务方面所达到的水平,以及信息抽取在公安领域案件文本中的应用情况,并提出了未来的研究趋势。

关键词:信息抽取 实体识别 事件抽取 案件文本

1.引言

在信息爆炸的今天,如何从海量的电子文档中快速有效地获得所需要的信息,成为了信息化时代亟待解决的问题,信息抽取便是在这种需求下应运而生的,以期把人们从大量、低效的文本阅读劳动中解放出来。

信息抽取是指从一段文本中抽取指定的一类信息并将其形成结构化的数据,填入一个数据库中供用户查询使用的过程。信息抽取已经成为自然语言处理研究中的一个热点,近年来在许多应用领域得以成功应用。

公安领域的文本信息主要有业务人员日常工作中记录下来的已经入库的半结构化案件信息,和一些没有入库的文档中的非结构化信息。面对日益增长的大量案件、涉案人员等信息数据,目前公安部门面临的问题是:侦查人员需要花费很多时间在阅读案件笔录上,如何将各类案件文本中的信息点分析出来,对涉案人员、案情信息等进行电子化管理,便于日后的查询与单位之间的信息共享;如何利用过往案件的信息,分析当前案情,挖掘线索,串并案件。信息抽取技术是解决这些问题的基础工作。

2.信息抽取的方法与技术概述

信息抽取系统的设计主要有知识工程方法和机器学习方法。

早期的信息抽取系统都是基于知识工程方法建立的,依靠人工编写抽取模式,使系统能处理特定知识领域的信息抽取问题。如CIRCUS系统、LIEP系统、PALKA系统、RAPIER系统等。规则本身的学习和提取成为信息抽取的关键,而信息抽取则退居为次要过程。这种方法要求编写抽取模式的知识工程师对该知识领域有深入的了解。而由人建立的规则很难保证具有整体的系统性和逻辑性,并且这些规则一般具有高度的领域相关性和较差的可移植性。因此,迫切需要寻找更加有效的方法来自动学习信息抽取的规则,这种形势使得机器学习在信息抽取系统中的应用研究显得尤为重要和迫切。

机器学习方法是利用机器学习技术让信息抽取系统通过训练文本来获得抽取模式,实现特定领域的信息抽取功能。任何对该知识领域比较熟悉的人都可以根据事先约定的规则来标记训练文本。利用这些训练文本训练后,系统能够处理没有标记的新的文本。BBN公司的SIFT系统,完全采用统计的方法,代表了在这个发展方向上跨出的重要一步。典型的机器学习方法有基于特征向量的机器学习方法,如支持向量机(SVM);有基于统计模型的机器学习方法,如隐马尔科夫模型(HMM)、最大熵模型(ME)和条件随机场模型(CRF);有基于核函数的机器学习方法,以及多种机器学习方法的集成等。现有研究成果表明,当多学习模型集成中的个体学习模型差异较大时,集成的效果会较好。

知识工程方法的设计初始阶段较容易,但是要实现较完善的规则库的过程往往比较耗时耗力。机器学习方法抽取规则的获取是通过学习自动获得的,但是该方法需要足够数量的训练数据,才能保证系统的抽取质量。所以,采取何种方法要视任务和资源而定,若训练语料容易获得,则倾向于机器学习的方法;若语言资源如词表等容易获得,则倾向于手工编写规则。

3.实体识别研究

命名实体识别的任务被定义为识别出文本中出现的专有名称和有意义的数量短语并加以归类。命名实体是文本中基本的信息元素,是正确理解文本的基础。狭义地讲,命名实体是指现实世界中的具体的或抽象的实体,如人、组织、公司、地点等,通常用唯一的标志符(专有名称)表示,如人名、组织名、公司名、地名等。广义地讲,命名实体还可以包含时间、数量表达式等。至于命名实体的确切含义,只能根据具体应用来确定。命名实体识别是信息抽取系统的一个基本而又重要的任务。

命名实体识别发展至今已经取得了很多成果。1987年开始由DARPA资助举办的MUC-6和MUC-7会议设立的命名实体专项评测大大推动了英语命名实体识别技术的发展,到1998年MUC最后一届会议时,不少系统都已经具备相当程度的大规模真实文本的处理能力,最好

语行为的理解能力相对较好;在面对面的交流情况下,智力落后儿童指示词的表达有障碍,而其对指示词的理解却与正常人相似;智障儿童在对话中更少引入新话题,智障成人经常对话题做一些没有意义的贡献,如用"哦"等来回应别人;在理想的环境下,发现儿童话轮转换的能力与正常人类似。总而言之,智障儿童的语用能力很多方面存在障碍,对其原因的探究还比较少,有待进一步研究,以便促进智障儿童的康复和治疗。

参考文献:

[1]引自吴昊雯,陈云英.智力落后儿童语用障碍研究新进展.中国特殊教育,2005,(6).

的成绩准确率和召回率达到了95%和92%。中文NE识别的难处在于其缺乏形式标志、分词错误会对其造成影响、内部常包含有常用字词以及词义模糊，需要更大量的研究工作。

命名实体识别任务要完成两个事情：一是找到文本中表达命名实体的词语，二是准确给出该命名实体的分类，其技术大多依赖于命名实体的类别。不同的类别所采用的识别技术也不一样。研究较多的几种类别是人名、地名、组织机构名、时间、数字。研究表明(张素香,2007)，不是一个模型能够完全解决所有的实体识别任务的，需要结合实体类型，采用不同的子模型识别能够极大地改善实体识别的性能。

中文命名实体的识别不光是信息抽取的基础，其研究同时也是分词、句法分析、问答系统、机器翻译等任务的基础，故对其研究，能从一定程度上对其他任务有所借鉴意义。

4.事件抽取研究

事件信息抽取(简称事件抽取)是信息抽取系统的另一个工作，是在命名实体识别基础之上实施的一个过程。其旨在利用计算机从文本中自动地抽取特定类型的事件及其事件要素，是信息抽取研究中最具挑战性的任务之一。

就前人研究情况来看，事件抽取主要有两种方法：模式匹配的方法和机器学习的方法。模式匹配的方法是指对于某类事件的识别和抽取是在一些模式的指导下进行的，采用各种模式匹配算法将待抽取的句子和已经抽出的模板匹配。例如Surdeanu和Harabagiu针对开放域的事件抽取系统FSA等。这种方法准确率较高，但往往依赖于具体领域，可移植性差。机器学习的方法把事件抽取任务看作分类问题，把主要精力放在分类器的构建和特征的发现、选择上。主要包括两个过程，即事件探测和事件元素识别。所谓事件元素，也就是平常所说的事件模板中的槽(Slot)，或事件的参与者(Participants)。

触发词为事件语句的锚定和事件类别的确定提供了很大的帮助。关于如何构建触发词集合，传统方法是将文本中每个词作为候选触发词，构建训练实例进行多元分类，但由于触发词只占候选触发词的一小部分，因此会引入大量的反例(赵妍妍,2008)。于江德(2007)对于“职务变动”类事件抽取的触发词表采用手工的方式构建，并借助于《现代汉语词典》和《同义词词林》，构建出的触发词表包含了136个职务变动类事件的触发词。赵妍妍(2007)使用哈工大信息检索研究室的《同义词词林(扩展版)》自动扩展种子触发词，通过查找过滤构建“种子触发词——事件类别”对照表，以便生成候选事件及其候选类别。

5.公安领域案件文本信息抽取研究概况

随着科技的进步，公安办公逐步实现了信息化，案件信息直接填入了相应的数据库中，因此该领域的工作大多集中于数据挖掘，即从已有数据中发现隐含的相似案件、犯罪趋势、犯罪特点等信息。但仍存在相当一部分的文档，或是侦查人员的案件笔录，或是网上的案件信息，以文本的形式存在，需要信息抽取技术从中抽取案件相关实体和事件，进而存入数据库中供后续的数据挖掘分析。

美国克莱蒙研究生院的Chih Hao Ku等人2008年报导正在开发一个自动的犯罪信息报导与调查访谈系统。该系统认为以往的格式化笔录由于种种原因会遗漏一些信息，故利用基于认知心理的访谈技术，唤起证人足够多的回忆信息，让其用自然语言记录案件情况，进而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。在信息抽取模块，采用了基于知识库和基于规则的方法。定义了“姓名、代词、时间、方式、武器、人物属性、场景、私人财物、颜色、身体部位、动作、事件、衣物”等实体类型。根据实体特点，针对性地利用一些如维基百科、网页博客、UCR官方信息、FrameNet等知识库资源，建立了一个有索引的词表，每个子表下设子类，如私人财物词条下设包、首饰、钱、电脑、电话等，如此产生了126个子表，分别应用于相应的规则构建。IE模块采用了Gate系统，包括：分词、索引、分句、词性标注、名词短语划分、正字校对、以及JAPE(Java Annotations Pattern Engine)规则构建等子模块。对于系统所产生的名词短语采用过滤算法，使提取的短语只与案件相关。(Chih Hao Ku et al.,2008(a); Chih Hao Ku et al.,2008(b); Alicia Iriberri et al.,2008)。另一个工作是美国亚利桑那州大学进行的一个基于神经网络的实体抽取系统。利用知识库、机器学习、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了识别和抽取。(Michael Chau et al.,2002; Hsin chun Chen et al.,2004)

国内在该领域对基于数据库的构建和数据挖掘技术研究的较多，对自然语言文本进行信息抽取研究的较少。乔春庚(2007)基于公安案件文本，对领域词汇的获取、命名实体的识别、实体关系的抽取等模块进行了研究。其搭建的分层的公安领域案件信息抽取系统，能够输出各层次的中间成果。徐亚娟(2008)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的功能。其在信息抽取阶段的算法主要思想是：根据分词结果得到的词性标注信息，通过扫描分词得到的结果串，去除一些无关的词性的词语，并结合专门的关键词库，完成信息的提取，最后得到结构化的文本信息，存入数据库中。

6.研究趋势

信息抽取是数据挖掘的第一步处理任务，若对案件文本进行了很好的信息抽取，不仅能够使业务人员免于阅读大量的案件，节省时间和人力，而且是后期的数据挖掘如串并相似案件、挖掘破案线索、归纳犯罪趋势等方面工作的良好基础。

现代信息抽取技术的研究，一方面，在努力地向投入应用发展，扩大抽取的文本类型的范围、扩大面向领域的范围，使科学技术能够真正地为人们生产生活提供方便，最大程度地解放劳动力；另一方面，在努力地探索如何加快其基础研究，使信息抽取技术实现革命性技术进步，使机器向高效自动处理任务迈进，尽量减轻研究者的劳动。这些，都需要学界人士的不断努力。

公安领域案件文本信息抽取研究综述

作者： 陈慧炜
作者单位： 南京师范大学,文学院,江苏,南京210097
刊名： 文教资料
英文刊名： DATA OF CULTURE AND EDUCATION
年，卷(期)： 2010(18)

本文读者也读过(10条)

1. 丁效.宋凡.秦兵.刘挺 音乐领域典型事件抽取方法研究[会议论文]-2010
2. 赵妍妍.秦兵.车万翔.刘挺 中文事件抽取技术研究7[会议论文]-2007
3. 于江德.肖新峰.樊孝忠 基于隐马尔可夫模型的中文文本事件信息抽取[会议论文]-2007
4. 潘霖 警备案事件信息提取与可视化方法研究[学位论文]2010
5. 毋菲.郑家恒 基于决策树的中文事件论元值的抽取[会议论文]-2009
6. 叶正.林鸿飞.苏绥.刘菁菁 基于支持向量机的人物属性抽取[会议论文]-2007
7. 丁效.宋凡.秦兵.刘挺.DING Xiao.SONG Fan.QIN Bing.LIU Ting 音乐领域典型事件抽取方法研究[期刊论文]-中文信息学报2011,25(2)
8. 赵妍妍.王啸吟.秦兵.车万翔.刘挺 中文事件抽取中事件类别的自动识别[会议论文]-2006
9. 赵健.王晓龙.关毅.徐志明.Zhao Jian.Wang Xiaolong.Guan Yi.Xu Zhiming 中文名实体识别:基于词触发对的条件随机域方法[期刊论文]-高技术通讯2006,16(8)
10. 刘辉 信息集成系统中面向领域的Web信息抽取研究[学位论文]2008